
THE THEORY OF EXTENDED TOPIC AND ITS APPLICATION IN INFORMATION RETRIEVAL

Ling Yin

A thesis submitted in partial fulfillment of the requirements of the
University of Brighton for the Degree of Doctor of Philosophy

September 2012

University of Brighton

ABSTRACT

This thesis analyses the structure of natural language queries to document repositories, with the aim of finding better methods for information retrieval. The exponential increase of information on the Web and in other large document repositories during recent decades motivates research on facilitating the process of finding relevant information to meet end users' information needs. A shared problem among several related research areas, such as information retrieval, text summarisation and question answering, is to derive concise textual expressions to describe what a document is about, to function as the bridge between queries and the document content. In current approaches, such textual expressions are typically generated by shallow features, for example, by simply selecting a few most-frequently-occurring key words. However, such approaches are inadequate to generate expressions that truly resemble user queries. The study of *what a document is about* is closely related to the widely discussed notion of *topic*, which is defined in many different ways in theoretical linguistics as well as in practical natural language processing research. We compare these different definitions and analyse how they differ from user queries. The main function of a query is that it defines which facts are relevant in some underlying knowledge base. We show that, to serve this purpose, queries are typically formulated by first (a) specifying a focused entity and then (b) defining a perspective from which the entity is approached. For example, in the query 'history of Britain', 'Britain' is the focused entity and 'history' is the perspective. Existing theories of topic often focus on (a) and leave out (b). We develop a theory of *extended topic* to formalise this distinction. We demonstrate the distinction in experiments with real life topic expressions, such as WH-questions and phrases describing plans of academic papers. The theory of extended topic could be applied to help various application areas, including knowledge organisation and generating titles, etc. We focus on applying the theory to the problem of information retrieval from a document repository. Currently typical information retrieval systems retrieve relevant documents to a query by counting numbers of key word matches between a document and the query. This approach is better suited to retrieving the focused entities than the perspectives. We aim to improve the performance of information retrieval by providing better support for perspectives. To do so, we further subdivide the perspectives into different types and present different approaches to addressing each type. We illustrate our approaches with three example perspectives: 'cause', 'procedure' and 'biography'. Experiments on retrieving causal, procedural and biographical questions achieve better results than the traditional key-word-matching-based approach.

TABLE OF CONTENT

ABSTRACT.....	2
TABLE OF CONTENT	3
TABLE OF TABLES.....	9
TABLE OF FIGURES.....	12
ACKNOWLEDGEMENT	13
AUTHOR’S DECLARATION	15
PUBLISHED WORK	16
CHAPTER 1	17
<i>Introduction</i>	<i>17</i>
1.1 Motivation	17
1.2 More Observations on Topics and User Queries	19
1.3 Automatic Topic Generation Approaches	21
1.4 Research Aims and Methodology	22
1.5 Overview of the Thesis	23
CHAPTER 2	25
<i>A Review of Theories of Topic and a Topic Classification Framework.....</i>	<i>25</i>
2.1 Introduction.....	25
2.2 Topic at the Sentence Level.....	26
2.2.1 Theoretical Work on Sentence Topic.....	26
2.2.1.1 Topic-Comment	26
2.2.1.2 Theme-Rheme	27
2.2.1.3 Given-New	29
2.2.1.4 Extensions of topic-comment	30
2.2.2 Summary of Sentence Topic	31
2.2.3 Thematic Analysis	32
2.2.3.1 Duchastel et al. (1992).....	32
2.2.3.2 Komagata (1999).....	33
2.3 Topic at the Discourse Level	33
2.3.1 Theoretical Work on Discourse Topic	34
2.3.1.1 Brown and Yule’s (1983) Topic Entity	34
2.3.1.2 van Dijk’s (1977) Discourse Topic	34
2.3.1.3 Hutchins’ (1977a) Aboutness	36
2.3.1.4 Brown and Yule’s (1983) Topic Framework	36
2.3.2 Automatic Topic Analysis at the Discourse Level	37
2.3.2.1 Norris’ (1998) Essence	38
2.3.2.2 Boguraev and Kennedy’s (1997) Topic Stamp.....	39

2.3.2.3 Rino and Scott's (1996) Gist	40
2.3.2.4 Marcu's (1997) Discourse Structure Approach	40
2.3.2.5 Script Based Approach	41
2.3.2.6 Latent Topic Model Approach	41
2.3.3 Summary of Discourse Topic and a Topic Classification Framework	42
2.4 Summary	46
CHAPTER 3	48
<i>Theory of Extended Topic</i>	48
3.1 Introduction	48
3.2 Theory of Extended Topic	49
3.3 Contexts in which Extended Topic are Explicitly Formulated	55
3.4 Outline of the Theory	58
CHAPTER 4	61
<i>Probing the Structure of Extended Topic</i>	61
4.1 Introduction	61
4.2 WH-Question Analysis	62
4.3 Experiment I	65
4.3.1 Introduction	65
4.3.2 Methodology	65
4.3.3 Data Preparation	65
4.3.4 Result	66
4.3.5 Conclusion	67
4.4 Experiment II	67
4.4.1 Introduction	67
4.4.2 Method	70
4.4.2.1 Data Acquisition	70
4.4.2.2 Term Generality Judgement by Human Subjects	73
4.4.2.3 Head Noun and Non-Head Noun Comparison	73
4.4.3 Result	74
4.4.3.1 Result of Term Distribution Experiment	74
4.4.3.2 Result of the General vs. Specific Scientific Term Experiment	75
4.4.3.3 Result of Testing the Proportion of General Scientific Term	75
4.4.4 Discussion	77
4.4.5 Experiment on Noun Phrases	78
4.4.5.1 Data collection	78
4.4.5.2 Result	78
4.4.6 Main Result	81
4.5 Experiment III	82
4.5.1 Introduction	82
4.5.2 Material and Method	82
4.5.2.1 Document Preparation	82

4.5.2.2 Signature Selection	83
4.5.2.3 Givenness Annotation.....	85
4.5.3 Result	86
4.5.4 Discussion	87
4.6 Summary	88
CHAPTER 5	89
<i>Applications of the Theory of Extended Topic.....</i>	<i>89</i>
5.1 Introduction.....	89
5.2 Knowledge and Text Indexing.....	89
5.3 Text Segmentation	90
5.4 Discourse Planning	93
5.5 Generating Indicative Topic Expressions for Passages	95
5.6 Automatic Retrieving Information from Documents.....	95
5.7 Summary	97
CHAPTER 6	98
<i>Apply Extended Topic to Information Retrieval</i>	<i>98</i>
6.1 Introduction.....	98
6.2 Basic IR Models.....	98
6.2.1 Introduction.....	98
6.2.2 Boolean Model	99
6.2.3 Vector Space Model.....	101
6.2.4 Probabilistic Model.....	105
6.2.5 Bayesian Inference Network Model	108
6.2.6 Language model.....	111
6.2.7 Limitation of IR Models for Retrieving Extended Topic	112
6.3 Query Expansion.....	117
6.3.1 Query Expansion Techniques.....	117
6.3.2 Limitation of Query Expansion.....	118
6.4 Apply Extended Topic to IR.....	119
6.5 Phrasal Retrieval for IR	122
6.6 TC and TC for IR	128
6.7 QA.....	130
6.7.1 Query Term Extraction & Information Retrieval Engine	133
6.7.2 Question Classification & IE	133
6.7.3 Answer Selection	134
6.7.4 Answering Non-Factoid Questions	135
6.8 Summary	136
CHAPTER 7	137
<i>Automatically Retrieving Relevant Documents for Causal Questions.....</i>	<i>137</i>
7.1 Introduction.....	137
7.2 Related Work.....	138

7.3 Methodology	139
7.4 Causal Indicators Preparation	140
7.5 Scoring Schema	142
7.6 Question Preparation	143
7.7 Corpus Preparation and Relevance Assessment	145
7.8 Experiments and Results	147
7.9 Discussion	149
7.10 Summary	150
CHAPTER 8	151
<i>Automatically Retrieving Relevant Documents for Procedural and Biographical Questions</i>	151
8.1 Introduction	151
8.2 Automatic Retrieving Procedures	152
8.2.1 Introduction	152
8.2.2 Related Work	153
8.2.3 Ranking Procedural Texts	153
8.2.3.1 Feature Selection and Document Representation	153
8.2.3.2 Corpus Preparation	156
8.2.3.3 Learning Method	156
8.2.3.4 Experiments	160
8.2.3.5 Discussion	162
8.2.4 Retrieving Relevant Documents for How-To Questions	164
8.2.4.1 Experiment Setup	164
8.2.4.2 IR Model	166
8.2.4.3 Result	166
8.2.5 Discussion	167
8.3 Automatic Retrieving Biographies	168
8.3.1 Introduction	168
8.3.2 Related Work	168
8.3.3 Text Categorisation for Identifying Biographical Documents.	169
8.3.3.1 Corpus Preparation	169
8.3.3.2 Document Preprocessing and Feature Selection	170
8.3.3.3 Document Representation	173
8.3.3.4 Classifiers	173
8.3.3.5 Experiments and Results	173
8.3.4 Automatically Retrieving Documents for Biographical Questions	174
8.3.4.1 Questions Preparation	174
8.3.4.2 System Architecture	175
8.3.4.3 IR Models	177
8.3.4.4 Result	177
8.3.5 Discussion	179
8.4 Summary	179

CHAPTER 9	181
<i>Conclusions and Future work</i>	181
9.1 Overview of Contributions	181
9.1.1 A Topic Classification Schema and the Nature of Indicativeness.....	182
9.1.2 Structure of Extended Topic	182
9.1.3 Relationship between Extended Topic Components and Discourse Constituencies	184
9.1.4 Two Modes of being a Topic.....	185
9.1.5 Applying Extended Topic to Information Retrieval.....	185
9.1.6 Other Contributions.....	186
9.2 Future Work	187
9.2.1 Model Other Generic Concepts to Improve Information Retrieval on Them	187
9.2.2 Study the Difference in Document Distribution between Generic Topic and Specific Topic	188
9.2.3 Relate Different Parts of Extended Topic to Different Discourse Constituencies	189
9.2.4 Other Future Work	190
REFERENCES	192
APPENDIX A.....	210
<i>WH-Question Analysis</i>	210
A.1 Questions from Medical Domain	210
A.2 Questions from TREC 2004.....	220
A.3 Questions from TREC 2007.....	228
APPENDIX B	230
<i>Generic Concept and Passage Selection</i>	230
B.1 Experiment Material.....	230
B.2 A Sample Questionnaire.....	248
APPENDIX C	258
<i>Study the Structure of Extended Topic</i>	258
C.1 Sample Topic Expressions.....	258
C.2 Algorithm for Extracting Head Nouns from a Noun Phrase.....	261
C.3 Evaluate the Head Noun Extraction Algorithm.....	264
C.4 Most Frequent Terms in Head Noun and non-Head Noun Lists.....	270
C.5 Questionnaire for Probing Term Generality	272
C.6 Inter-Rater Agreement Test	276
C.7 Term Classification Result	278
C.8 Lists of Most Frequent General Terms at the Head Noun Position	280
APPENDIX D.....	282
<i>Verify the Relation between Different Parts of Topic and Different Discourse Constituencies</i>	282
D.1 Experiment Material	282
Group 1	282
Group 2	284

Group 3	286
D.2 Topic Signatures	288
APPENDIX E	291
<i>Comparison between Generic Topic and Specific Topic</i>	291
E.1 Concept: Biography.....	291
E.1.1 Biography as a Generic Topic	291
E.1.2 Biography as Part of a Specific Topic	294
E.2 Concept: Anatomy	297
E.2.1 Anatomy as a Generic Topic.....	297
E.2.2 Anatomy as Part of a Specific Topic	300
APPENDIX F	303
<i>Answering Causal Questions</i>	303
F.1 Causal Indicators and Usage Examples	303
F.2 Patterns of Causal Indicators	306
F.3 Sample Experiment Text	307
APPENDIX G	310
<i>Answering Procedural Questions</i>	310
G.1 Procedural Feature Set.....	310
G.2 Sample Experiment Text	312
APPENDIX H.....	316
<i>Answering Biographical Questions</i>	316
H.1 Sample Biographies	316
H.2 An Example Biographical Feature Set.....	321

TABLE OF TABLES

Table 4-1. A classification of generic concepts	63
Table 4-2. The selected generic concepts.....	66
Table 4-3. Compare the level of agreements between set 1 and set 2.....	67
Table 4-4. Compare the level of agreements between set 2 and a random distribution	67
Table 4-5. Evaluate the level of agreements between subjects' choices in set 1 and the expected result ..	67
Table 4-6. Number of collected topic expressions	71
Table 4-7. Number of terms and term frequencies.....	74
Table 4-8. Term distributions	74
Table 4-9. Agreement between two subjects based on kappa statistic	75
Table 4-10. Statistics of the representative terms	76
Table 4-11. Difference of the proportions of scientific-research-general terms.....	76
Table 4-12. Chi-square statistics.....	76
Table 4-13. Comparison between proportions of scientific-research-general terms	77
Table 4-14. Number of collected noun phrases	78
Table 4-15. Number of terms and term frequencies in noun phrases	79
Table 4-16. Term distributions in noun phrases.....	79
Table 4-17. Statistics of the representative terms in noun phrases.....	80
Table 4-18. Difference of the proportions of scientific-research-general terms in noun phrases	80
Table 4-19. Chi-square statistics for noun phrases.....	80
Table 4-20. A list of general terms and the frequencies in both 'CL' and 'Physics'	81
Table 4-21. Groups of generic topics and associated specific topics	83
Table 4-22. Documents in group1 and their topics	83
Table 4-23. The numbers of occurrences of topic signatures in different categories—group 1.....	86
Table 4-24. The numbers of occurrences of topic signatures in different categories—group 2.....	86
Table 4-25. The numbers of occurrences of topic signatures in different categories—group 3.....	87
Table 6-1. Various schemata to calculate vector distances.....	104
Table 7-1. Patterns of linkage phrase	141

Table 7-2. Question collection.....	145
Table 7-3. Average precisions of four different methods using all the questions.....	147
Table 7-4. The significance of differences among four methods using all 21 questions.....	148
Table 7-5. Average precisions of five different methods using part of the questions	148
Table 7-6. The significance of differences among five methods using part of the questions	149
Table 8-1. Sample cue phrases and matching patterns	154
Table 8-2. The 2*2 contingency table for a feature cooccurrence pattern.....	155
Table 8-3. Ranking results using individual features	161
Table 8-4. Ranking results using feature cooccurrence patterns	162
Table 8-5. Ranking results using selected individual features.....	163
Table 8-6. Procedural question set	164
Table 8-7. Results of different systems.	167
Table 8-8. Five meta-tags	170
Table 8-9. The 2*2 contingency table for measuring the distinctiveness of a pattern	171
Table 8-10. The 2*2 contingency table for calculating Phi2.....	172
Table 8-11. The 2*2 contingency table for calculating Phi3.....	172
Table 8-12. Text categorisation result.....	174
Table 8-13. Biographical question set	175
Table 8-14. MAPs of different systems	178
Table 8-15. Pairwise t-tests results (BM25).....	178
Table 8-16. Pairwise t-tests results (PL2)	179
Table C-1. Top 10 percent head noun terms and frequencies in 'CL+Describe'	270
Table C-2. Top 5 percent non-head noun terms and frequencies in 'CL+Describe'	270
Table C-3. Top 10 percent head noun terms and frequencies in 'CL+Present'	271
Table C-4. Top 5 percent non-head noun terms and frequencies in 'CL+Present'	271
Table C-5. Top 10 percent head noun terms and frequencies in 'Physics+Describe'.....	271
Table C-6. Top 5 percent non-head noun terms and frequencies in 'Physics+Describe'	271
Table C-7. Top 10 percent head noun terms and frequencies in 'Physics+Present'	271

Table C-8. Top 5 percent non-head noun terms and frequency in 'Physics+Present'	272
Table C-9. Term classification result.....	279
Table C-10. Most frequent general terms in 'CL+Describe'	280
Table C-11. Most frequent general terms in 'CL+Present'	280
Table C-12. Most frequent general terms in 'Physics+Describe'	280
Table C-13. Most frequent general terms in 'Physics+Present'	281
Table F-1. Causal indicators and usage examples	305
Table F-2. Patterns of causal indicators.....	307
Table G-1. Full procedural feature set.....	311
Table G-2. Selected distinctive individual feature set	312
Table H-1. Example biographical feature set.....	321

TABLE OF FIGURES

Figure 3-1. Selecting from a knowledge base.....	50
Figure 3-2. An example knowledge base.....	51
Figure 3-3. Mozart family tree.....	52
Figure 3-4. Composition of extended topic.....	53
Figure 4-1. Narrow vs. wide term distributions.....	69
Figure 5-1. A topic tree for an article about coronary artery disease (Kan et al., 2001)	94
Figure 6-1. Boolean query expression examples.....	99
Figure 6-2. A Bayesian network model example	109
Figure 6-3. A simplified Bayesian network model	109
Figure 6-4. A general architecture of QA systems.....	131
Figure 8-1. An illustration of the problems in using the Naive Bayes classification algorithm	158
Figure 8-2. An illustration of the Adapted Naive Bayes classification algorithm	159
Figure 8-3. Ranking results using individual features: 1 refers to Adapted Naive Bayes, 2 refers to Naive Bayes, 3 refers to ME, 4 refers to ADTree and 5 refers to Linear Regression	161
Figure 8-4. Ranking results using feature cooccurrence patterns: 1 refers to Adapted Naive Bayes, 2 refers to Naive Bayes, 3 refers to ME, 4 refers to ADTree and 5 refers to Linear Regression	162
Figure 8-5. Ranking results using selected individual features: 1 refers to Adapted Naive Bayes, 2 refers to Naive Bayes, 3 refers to ME, 4 refers to ADTree and 5 refers to Linear Regression	163
Figure 8-6. A two-stage architecture.....	165
Figure 8-7. An alternative architecture using query expansion	165
Figure 8-8. MAPs of different systems: 1 refers to using BM25 as the IR model, 2 refers to using PL2 as the IR model.....	167
Figure 8-9. Text categorisation result: 1 refers to Naive Bayes, 2 refers to SVM, 3 refers to ADTree	173
Figure 8-10. The system that applies the query expansion technique.....	175
Figure 8-11. The two-stage architecture	176
Figure 8-12. MAPs of different systems: 1 refers to BM25, 2 refers to PL2	178

ACKNOWLEDGEMENTS

I am very grateful to my supervisors, Dr. Richard Power and Dr. Roger Evans, for their countless discussions with me, detailed comments and revisions on my drafts, and guidance and encouragement. They influenced me with the spirit of scientific research and helped me improve my logical thinking. When I was deeply confused about the direction I was taking with my research, they listened to me patiently, shared their thoughts and profound advice. They provided me with a strong support to pursue my research.

I appreciate Professor Donia Scott for providing me with the opportunity to study in England and guiding me in the first three years of my PhD program. She taught me to be more determined and persistent when facing difficulties. I was also encouraged by her to attend summer school and conferences, which helped me build connections with people working in related research areas.

Dr. Tianfang Yao was my supervisor during my graduate study. He brought me into the field of computational linguistics and also opened the door for me to experience Europe. My life and career would have gone into a very different path without his help.

I am truly thankful to my examiners, Dr. Udo Kruschwitz and Dr. Lynne Cahill, for their useful criticisms, detailed revision suggestions, and for introducing me to a list of further readings.

I have fond memories of a nurturing companionship with Dr. Marina Santini. We started our PhD studies at a similar time and spent the first three years together. I truly admire her courage to pursue her passion in web genre detection. In addition to our academic studies, we often toured castles, museums, and many ancient monuments such as Stonehenge, where she introduced the European art and culture to me. We also enjoyed the different aspects of Brighton, including nature and the downtown life, as we were walking at the seafront, driving to the little pubs on the countryside, and sitting at concerts and operas captivated by the exquisite performance.

I would also like to thank all the other colleagues in the former information technology research institute (ITRI) and in the current natural language technology group (NLTG) under the School of Computing, Engineering and Mathematics research. We have different cultural backgrounds but still understand each other very well. They are, roughly in chronological order, Kees van Deemter, Lynne Cahill, David Tugwell, Gabriela Cavaglià, Jon Herring, Ivandré Paraboni, Daniel Paiva, Adam Kilgarriff, Anja Belz, Paul Piwek, Amy Neale, Leigh Dodds,

Martyn Haddock, Catalina Hallett, Sebastian Varges, Petra Tank, Albert Gatt, Thapelo Otlogetswe, Jason Teeple, Norton Trevisan Roman, Aylin Kocha and Michel Génèreux.

My memory of the time spent with all the housemates in 15th Nyetimber Hill will never diminish. Most of them are international students and their company made me feel at home. I would especially like to thank Freida Ibaduni M'Cormack for organising the big parties, Phillip Doughty for all the help that he offered to me, Martin Krawczynszyn for playing silly games with me, and Elya Othman for sharing his sagas and teaching me how to smoke.

I also got to know a lot of Chinese friends during my stay in England, including Chiho (Henry) Li, Xinglong Wang, Haihua (Carrie) Zhang, Guang (Grace) Shi and Han Jiang. We often gathered to celebrate traditional Chinese festivals and try out new things with each other.

I treasure all the memorable moments I spent with my ex-boyfriend, Yuehua Chen. I will never forget the days when we were flying to see each other, learning to take care of each other, and happily exploring Europe together. I would not have been able to finish my studies without the spiritual support that he provided me.

Thanks to my family, including my father, Yubo Yin, my mother, Dianfang Wang, and my sister, Bingsheng Yin, for always being there and supporting me from a remote country.

Thanks to England for the generosity and the spirit of democracy and justice.

AUTHOR'S DECLARATION

I declare that the research contained in this thesis, unless otherwise formally indicated within the text, is the original work of the author. The thesis has not been previously submitted to this or any other university for a degree, and does not incorporate any material already submitted for a degree.

Signed

Dated

PUBLISHED WORK

Yin, L. (2004). Topic analysis and answering procedural questions. ITRI Technical Report 04-14, ITRI, University of Brighton. <<http://www.itri.brighton.ac.uk/techindex.html>> [accessed June 1st, 2005]

Yin, L. and R. Power (2005). Investigating the structure of topic expressions: a corpus-based approach. In *Proceedings from the Corpus Linguistics Conference Series*, Vol.1, No.1, University of Birmingham, Birmingham.

Yin, L. (2006). A Two-Stage Approach to Retrieve Answers for How-To Questions. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*. Student session, Trento, Italy

Yin, L. and R. Power (2006). Adapting the Naive Bayes classifier to rank procedural texts. In *Proceedings of the 28th European Conference on IR Research (ECIR 2006)*.

CHAPTER 1

INTRODUCTION

1.1 MOTIVATION

In the last three decades, the proliferation of personal computers and the emergence of the Internet have dramatically changed the process of authoring and accessing information. Publishing information on the Internet has become a routine everyday activity. The rapid growth of the Web and its low technical requirement make it the number one information resource. However, the ever-increasing amount of information freely available online or in large data repositories has also made information accessing more difficult. This has motivated the research on information organisation and on ways to automate the process of locating relevant information.

Extensive studies have been carried out in the area of information retrieval (IR), text categorisation (TC), information extraction (IE), question answering (QA) and text summarisation. Information retrieval studies automatically retrieving documents containing relevant information to a user query (typically a list of key words or a natural language question) from the Web or a large document repository. Text categorisation studies automatically classifying documents into several predefined categories (based on their topics or genres). The techniques used in a TC system are very similar to an IR system since the latter could be seen as a binary TC system which classifies documents into a relevant set and an irrelevant set. Compared to IR and TC, information extraction and question answering systems aim to locate the information in a more accurate way. An IE system analyses documents and automatically extracts relevant text snippets or concepts to fill the informational slots in some predefined topic schemata. QA could be seen as the combination of IR and IE. It automatically extracts answers to a user query from relevant documents retrieved on the Web or in a large document repository. Most current QA systems adopt a two-stage approach: first using IR technology to find the relevant documents; then further using IE techniques to pinpoint the answers. Instead of finding relevant information for a given topic/query/genre, automatic text summarisation takes a different approach: it helps users to find relevant information by generating “a condensed representation of the content of an information source in a manner sensitive to the needs of the user and task” (Mani and Maybury, 1999).

A shared issue among these research areas is the study of high-level topic expressions that are representative of the discourse content. We will give a formal definition of what topic expression means later. An IR system typically approaches such a representation by simply extracting all the keywords in the document weighted by their frequency of occurrences. Document retrieval would then simply be the match of keywords in user query against those in document. Taking passage 1.1 below as an example, which is a short biography of Steve Jobs. When user queries about ‘Steve Jobs’, an IR system would be able to retrieve it because the keyword in the query, i.e., ‘Steve’ and ‘Jobs’, also appear in the passage; however, when user queries about ‘biography of Steve Jobs’, the system would not be able to find it since the word ‘biography’ is not explicitly mentioned in the passage. We see that the concept ‘Steve Jobs’ and ‘biography’ differ in that the former refers to an entity at the centre of the discussion and the latter describes the kind of information being discussed about this entity. Thus, the topic expression ‘biography of Steve Jobs’ contains two different parts with different roles. We will show later that such a structure could be seen in many human-formulated topic expressions. In recent years, the surge of the Web attracts more research on Web search. Web pages contain much more structured information than traditional digital documents, including the link structure, the anchor texts, the user queries and clicks on the Web search results, etc. The structured information provides a good summary of the content in the Web page which will potentially help information retrieval. Detailed discussions on Web search is beyond the scope of this thesis.

[1.1] **Steve Jobs** was a college dropout when he teamed up with **Steve Wozniak** in 1976 to sell personal computers assembled in **Jobs'** garage. That was the beginning of Apple Computers, which revolutionized the computing industry and made **Jobs** a multimillionaire before he was 30 years old. He was forced out of the company in 1985 and started the NeXT Corporation, but returned to his old company in 1996 when Apple bought NeXT. **Jobs** soon became Apple's chief executive officer and sparked a resurgence in the company with products like the colorful iMac computer and the iPod music player. **Jobs** is also the CEO of Pixar, the animation company responsible for movies like Toy Story and Monsters, Inc. Pixar was purchased by the Walt Disney Company in 2006 for \$7.4 billion in stock; the deal made **Jobs** the largest individual shareholder of Disney stock. **Jobs** was diagnosed with pancreatic

cancer in 2003 and had surgery in July of 2004, and was criticized by some for not disclosing his illness to stockholders until after the fact.¹

The above-mentioned topic structure and the difference between the two parts are not defined in the literature of topic and automatic topic generation. In fact, most existing theories just focus on the central entities. It is the aim of the thesis to understand the structure of topic expressions and develop a theory of *extended topic* to formalise the distinction. We will also investigate ways in which the theory could be applied to improve various applications. In this thesis, we emphasize that topic expressions must resemble user queries to better fit for the requirement of the application area, i.e., to help users finding relevant information.

This chapter is organised as follows: section 1.2 provides more examples to demonstrate the structure of topic expressions; section 1.3 discusses in more details about the problem in existing automatic topic generation systems; section 1.4 presents the concrete research aims and methodologies; section 1.5 provides an overview of the whole thesis.

1.2 MORE OBSERVATIONS ON TOPICS AND USER QUERIES

Above we showed that some topic expressions contain two different parts with different roles. Here we will use a few examples to further illustrate the structure.

A simple and intuitive definition of topic is that *topic is what a sentence/discourse talks about*. For example, for sentence [1.2], we might say that it talks about lilies. This approach of simply taking part of the sentence does not give us much hint of what the rest is about. Another answer would be that it is about ‘the colour of lilies’. The concept ‘colour’ well characterises the comment made about lilies. We can easily abstract the concept ‘colour’ from the detailed comment ‘white’. The same process applies to passage [1.3]: at first sight, we identify some entities in the centre of the discussion, such as ‘jaw’, ‘mandible’ and ‘muscles’; a better characterisation of the content of the discourse would be ‘the anatomy of the jaw’. The concept ‘anatomy’, which indicates the spatial relationships between parts, captures which aspect of jaw is being talked about. There might be different types of topic. In fact, there are many different theories of topic in theoretical linguistic studies. As mentioned above, in this thesis, we emphasizes that topic expressions should resemble user queries. In each of the two

¹ Cited from <http://www.answers.com/topic/steve-jobs>

examples given above, the second topic expression meets this requirement better than the first one. We could imagine that sentence [1.2] is the answer to question ‘what is the colour of lilies?’ and passage [1.3] is the answer to question ‘what is the anatomy of the jaw?’; in comparison, rarely would a user generally ask about ‘lilies’ or about ‘jaw’.

[1.2] Lilies are white.

[1.3] The jaw is the lowest and only mobile bone of the face, also known as the mandible. The mandible is U-shaped as seen from above and bears the lower teeth on its upper surface. It is connected to the base of the skull at the temporomandibular joints, which can be felt in the cheek just in front of the earlobe. Powerful muscles, arising from the temple on either side, attach to the jaw for movements needed in chewing and biting; other muscles allow side-to-side and downward movement.

The above-described topic structure, with a concrete entity and an abstract concept that denotes from which *perspective* it is approached, could be seen in many human-formulated topic expressions. Passage [1.4] is excerpted from the Medline². The term ‘function of calcyclin’ in sentence b and the term ‘the functional role of calcyclin’ in sentence c indicate the topic of the second clause in sentence b. The referring expression ‘these observations’³ (of calcyclin) in sentence g indicates the topic of sentence d, e and f.

[1.4] a. Calcyclin (S100A6) is a member of the S100A family of calcium binding proteins. b. While the precise **function of calcyclin** is unknown, calcyclin expression is associated with cell proliferation and calcyclin is expressed in several types of cancer phenotypes. c. In the present study, the **functional role of calcyclin** was further elucidated in pulmonary fibroblasts. d. Antisense S100A6 RNA expression inhibited serum and mechanical strain-induced fibroblast proliferation. e. This attenuated proliferative response was accompanied by a flattened, spread cell morphology, and disruption of tropomyosin labeled microfilaments. f. Changes in cytoskeletal organization did not correspond with a decrease in tropomyosin levels. g. **These observations** suggest a role for calcyclin in modulating calcium

² Refer to <http://www.ncbi.nlm.nih.gov/PubMed/>

³ The complete form is ‘these observations of Calcyclin’.

dependent signaling events that regulate progression through the cell cycle.

1.3 AUTOMATIC TOPIC GENERATION APPROACHES

In the research areas that study how to automate the process of finding relevant information, the distinction between the two parts within a topic expression is not explicitly addressed. Most systems use shallow features such as positional information and word frequencies to generate topic expressions; such approaches could only effectively detect the central entities being talked about in the discourse.

Boguraev and Kennedy (1997) generate a concise representation of discourse content by extracting the most *salient* elements from discourses. They note that “objects at the centre of discussion have a high degree of salience; objects at the periphery have a correspondingly lower degree of salience”. Salient elements are identified by term frequencies and syntactic positions (e.g., the subject of a sentence is more salient than the object of it). For passage [1.5], the system generates “desktop machines” and “operating systems”.

[1.5] Lots of “ifs,” but you can't accuse Amelio of lacking vision. Today's desktop machines, he says, are ill-equipped to handle the coming power of the Internet. Tomorrow's machines must accommodate rivers of data, multimedia and multitasking (juggling several tasks simultaneously). We're past the point of upgrading, he says. Time to scrap your operating system and start over. The operating system is the software that controls how your computer's parts (memory, disk drives, screen) interact with applications like games and Web browsers. Once you've done that, buy new applications to go with the reengineered operating system.

Such topic expressions capture objects in the centre of the discussion but apparently leave out what is being said about them. In comparison, the authors of the paper describe the content of the above passage as “the future of desktop computing”. Similarly, for a few other passages, the system generates “Apple” and “Microsoft” in contrast to “the relation between Apple and Microsoft” given by the authors, “Gilbert Amelio” and “new operating system” in contrast to “Amelio's background” and “plans for Apple's operating system”. The terms ‘relation’, ‘future’, ‘background’ and ‘plan’ are abstracted from the details and provide a good characterisation of them.

Above in section 1.1 we mentioned that typical key-word-matching-based IR approach is effective in retrieving documents about 'Steve Jobs' but not so for 'biography of Steve Jobs'. It also appears that some IR systems only use part of user queries to match retrieve relevant documents. Lioma and Ounis (2006) study how to select effective terms from queries to retrieve documents. For an example query "a relevant document will focus on the causes of the lack of integration in a significant way; that is, the mere mention of immigration difficulties is not relevant. Documents that discuss immigration problems unrelated to Germany are also not relevant", they extract 'lack of integration' and 'immigration difficulties' and remove word 'causes'.

We also mentioned in section 1.1 that QA systems typically contain two stages: first use IR technology to retrieve relevant documents; then use IE technology to pinpoint the answers. In Saxena et al. (2007), they classify questions into different semantic types, including LOCATION, ORGANIZATION and JOB, etc. To find the answers, they first use the keywords in the original question to match retrieve relevant passages; then match each sentence against a few lexical patterns predefined for the question type. The two-stage framework, with each stage tackling one part of a question, also supports the observation that a topic expression contains two different parts.

1.4 RESEARCH AIMS AND METHODOLOGY

Above we show that topics typically contain two parts and one part is often ignored in existing literature of automatic topic generation. The general aim of the thesis is to study topic expressions to formalise the topic structure and to apply the insights drawn from the study to improve practical problems, in particular, to facilitate retrieving relevant information. In addition to automatic topic generation, the notion of topic is widely discussed in theoretical linguistic studies. It is necessary to review these work to understand whether our observation of topics is addressed there and to clarify the difference among different types of topic. The concrete research aims are formulated as below.

Research aim 1: develop a topic classification framework to compare existing theories of topic and to help us define the difference between existing theories and user queries to document repositories.

We will review both the theoretical work and the practical work on topic. We will develop a framework to classify existing theories of topic taking consideration of the function, the syntactic form and the content of topics. This topic classification framework will help clarify

the difference among existing theories of topic. The function of topic expressions might also help explain their structure.

Research aim 2: analyse the structure of user queries and study how different components in a query are derived from the discourse constituencies.

The examples given in above sections indicate that there are different parts in a topic, one part (concrete entities) could be easily identified from the discourse and the other part (the perspectives) is abstracted from the detailed content in the discourse. Existing theories of topic focus on the concrete entities and omit the perspectives. We will develop a theory of extended topic to formalise the distinction and will experiment on real topic expressions to verify the theory. Furthermore, we will study the distinction by investigating the relations of the different parts in topic to different discourse elements. Grammars such as the information structure theory provide base to classify elements in sentences/discourses into different types. This investigation will also provide insights in the ways to automatically generating perspectives from discourse elements.

To improve the knowledge of the perspectives, we will also extract the perspectives in real topic expressions and classify them into different types.

Research aim 3: study ways to apply the theory of extended topic to various applied contexts, in particular, to improve retrieval of relevant documents.

In section 1.2, we noted that the perspectives in user queries are often removed when extracting keywords to match retrieve relevant documents. Since the perspectives are also very important in defining relevant information, we will propose concrete approaches to improve retrieving the perspectives. The analyses on the different types of perspectives and how the perspectives are derived from the discourse elements will provide insights in this direction.

1.5 OVERVIEW OF THE THESIS

The thesis is composed of two parts. The first part includes chapter 2 to chapter 5. It develops a theory of extended topic based on the analysis of the function and the structure of user queries. In chapter 2, we will review theoretical and practical studies on topic and develop a topic classification framework. In chapter 3, we develop a theory of extended topic to formalise the structure of user queries and define how different parts of extended topic are derived from the discourse. In chapter 4, we design a series of experiments on real topic

Chapter 1 – Introduction

expressions to verify the theory of extended topic. Chapter 5 provides ways to apply the theory to various application areas including information retrieval and knowledge organisation, etc. The second part includes chapter 6 to chapter 8. It focuses on applying the theory to improve IR. Chapter 6 reviews the state of the art of IR technologies, analysing in details why they do not fit for retrieving the perspectives. It also proposes different approaches to improve retrieving different types of perspectives. Chapter 7 and chapter 8 present IR experiments to verify the proposed approaches. Chapter 9 summarises all the findings in the thesis and provides future research directions.

CHAPTER 2

A REVIEW OF THEORIES OF TOPIC AND A TOPIC CLASSIFICATION FRAMEWORK

2.1 INTRODUCTION

In the last chapter, we mentioned that many existing topic generation systems select the most prominent terms from a discourse to form a topic expression, where prominence is defined by term frequency or its syntactic position; topics generated by this approach lack an important part compared to human-generated topics.

The notion of topic is widely discussed in the study of linguistics and computational linguistics. We will review different theories of topic in this chapter, including theoretical linguistic work and systems that deal with practical computational problems, with the aim being to better understand the difference between different types of topic, the structure of topic and how topic is derived from a discourse. Specifically, we will try to answer the following questions:

- a) under what contexts is the notion of topic proposed?
- b) what functions does topic play?
- c) what elements in a sentence/discourse are identified as the topic?
- d) what is the composition of a topic?

The notion of topic is further divided into two levels, i.e., topic of a sentence/clause and topic of a discourse. We will review theories of topic at the sentence/clause level in section 2.2 and will then review theories of topic at the discourse level in section 2.3. We will show that different theories of topic are developed for the purpose of analysing different problems, such as to define sentence structure or to analyse how a discourse is organised. As mentioned in chapter one, the practical aim of the thesis is to facilitate the process of finding relevant information for user queries; thus our focus is on analysing the role of topic in information (knowledge) indexing and sharing. At the end of each section, we will extract the points relevant to this focus from different theories. In section 2.3.3, we develop a topic classification framework which groups existing theories of topic into two types: *indicative topic* and *informative topic*. We compare between these two types of topic in terms of the function, the

structure and the syntactic form and conclude that indicative topic align better with the practical aim mentioned above. Section 2.4 recapitulates some major points of this chapter.

2.2 TOPIC AT THE SENTENCE LEVEL

In this section, we will first review different theories of sentence topic and then introduce a few studies on automatic topic analysis.

2.2.1 THEORETICAL WORK ON SENTENCE TOPIC

At the sentence level, topic is usually treated as a grammatical term which identifies a constituent in the structure of a sentence (Brown and Yule, 1983, p. 70). Specifically, a distinction is drawn between topic and the notion of comment. The topic-comment system provides an account of sentence structure. In addition to the topic-comment distinction, there are two other theories which also provide structural analysis on sentences, i.e., the theme-rheme system and the given-new system. These theories are closely related. In fact, different scholars may name the systems differently, for example, some would use given-new to refer to what others usually call theme-rheme. Below we will introduce the definition of all three systems. It is worth noting that most studies on sentence topic focus on conversations or speeches, but the concept of sentence topic and its definition could well apply to written work.

2.2.1.1 TOPIC-COMMENT

The notion of topic is rooted in some early work by Mathesius (1915) and Hockett (1958). Brown and Yule (1983, p. 70) study Hockett's work and define the general structure of sentence as *"the speaker announces a topic and then says something about it. ... In English and the familiar languages of Europe, topics are usually also subjects and comments are predicates"*. Sentence [2.1] and [2.2] are given by Hockett (1958). In sentence [2.1], John is the topic and the rest is the comment. However, topic is not necessarily the subject. For example, in sentence [2.2], topic is a pre-posed object.

[2.1] John | ran away.

[2.2] That new book by Thomas Guernsey | I haven't read yet.

Early work such as Mathesius (1915) and Hockett (1958) define topic as a grammatical notion identifying a sentential constituent, later on a pragmatic notion of topic has also been introduced, such as that given by Gundel (1988).

Topic Definition:

“An entity, E, is the topic of a sentence, S, iff in using S the speaker intends to increase the addressee’s knowledge about, request information about, or otherwise get the addressee to act with respect to E.”

Comment Definition:

“A predication, P, is the comment of a sentence, S, iff, in using S the speaker intends P to be addressed relative to the topic of S”.

The pragmatic notion of topic defines *topichood* based on a notion of *aboutness*. Vallduví (1990, p. 40 – 41) discusses the early work by Gundel (1974) and Reinhart (1982), which have set a number of tests to provide an operational tool to identify the topic of a sentence, including the ‘as-for’ test, the ‘what-about’ test and the ‘said-about’ test, etc. The ‘what-about’ test, for example, establishes that an NP is the topic of a sentence if this sentence can answer the question ‘what about x?’. Here we will not talk in details about the other two tests.

Therefore, in contrast to the grammatical definition of topic, which considers the element taking the initial position in a sentence as the topic, Gundel (1988) allows any referential phrase in a sentence to be the topic of that sentence, depending on the interpretation intended. Furthermore, she notes that a pragmatic topic needs not to be an overt expression in the sentence at all, and the topic of a sentence depends on the context under which the sentence is interpreted. Below is an example that she gives.

[2.3] Marcos resigned

The focus (or comment) of this sentence, would be “the primary stressed constituent *resigned* in a context where the topic is Marcos or what Marcos did and the comment is that he resigned”; alternatively, the comment could be “the whole sentence under the interpretation where the topic is some entity not overtly expressed”, for instance, “the political situation in the Philippines or what happened there at a particular time”. (Gundel, 1988, p. 211)

2.2.1.2 THEME-RHEME

The notions theme and rheme, according to Vallduví (1990, p. 36), have its root in the notions of *thema* and *rhema* introduced by Ammann (1928). However, some latter work shows different interpretations for the theme-rheme system. As Vallduví (1990, p. 36) notes, there are at least two themes, one being introduced by Halliday (1967) and another being

introduced by Firbas (1964, 1971) and Daneš (1968). Halliday (1967) uses the notion theme in two senses. In its broad sense, theme, together with *transitivity* and *mood*, constitutes three main areas of syntactic choices for English clauses. Specifically, he remarks that “theme is concerned with the information structure of the clause; with the status of the elements not as participants in extralinguistic processes but as components of a message; with the relation of what is being said to what has gone before in the discourse, and its internal organization into an act of communication”. There are a set of different functions defined under the area of the theme system. In its narrow sense, the notion of theme refers to one particular function in this set. Halliday (1967) defines the given-new system to address other functions. Firbas (1964) and Daneš (1968) define theme-rheme under a theoretical framework of communicative dynamism. Their definition of theme-rheme is similar to the given-new system given by Halliday. Below we will focus on the narrow sense of the theme-rheme system defined by Halliday and will introduce Firbas’s definition under the topic given-new.

Halliday (1967) asserts that, when ordering elements into a clause, the speaker marks what he wants to talk about at the beginning. Therefore, the first element in a sentence, which delivers what the speaker intends to talk about, is identified as the *theme*; *rheme* refers to all the elements following theme. *Theme* establishes the point of departure for the clause as a message.

Brown and Yule (1983, p. 127) give examples (as in [2.4]) which effectively show that several different sentence forms, which convey exactly the same propositional meaning but differ in their elements ordering, are only appropriate in different context. To be more concrete, they seem to be appropriate answers for different questions.

- [2.4] a. John kissed Mary.
- b. Mary, John kissed her.
- c. Who John has kissed is Mary.

For example, sentence *a* seems to be an appropriate answer for question ‘what did John do?’, but is less appropriate as an answer to question ‘what happened to Mary?’, for which *b* seems to be more appropriate. These suggest that the ordering sequence does make a difference. Sentence *c* seems to presume that the hearer knows that John has kissed somebody. Therefore, when selecting the theme, the speaker also has to take the hearer’s state of knowledge (what he knows) and the hearer’s expectation (what he understands is being talked about) into account.

Since the theme of a sentence takes the initial position, the theme of a declarative sentence is, most of the time, the grammatical subject. Until now, we have seen that theme is quite similar to the notion of topic, in that they both often overlap with grammatical subject and they are both related to a notion of aboutness.

While Halliday (1967) defines theme as what the speaker wish to talk about now in a given clause, independent of what has been talked about before, Brown and Yule (1983) believe that theme has its role at the discourse level. Specifically, they observe that many sentences in a discourse put the same referent at the initial position of the sentence. The referent becomes very prominent in the discourse and represents what the author wishes to talk about. They call this phenomenon thematisation and believe it is one basic organisation principle of a discourse to place the main referent to the subject position in sentences within it. Thus, the theme of a sentence also plays the function of “connecting back and linking in to the previous discourse, maintaining a coherent point of view”. We will talk more about thematisation in section 2.3 when introducing theories of topic at the discourse level.

2.2.1.3 GIVEN-NEW

The given-new theory is developed under the study of information structure, which studies how information is organised into small units at the phrase and clause level. As noted before, the given-new system is proposed by Halliday (1967) but has a deep root in Firbas (1964, 1971) and Danes (1968), who use terms theme and rheme instead. Halliday (1967, p. 199-244) notes that any text in spoken English is organised into what may be called “information units”. “In each information unit, there is information focus which selects a certain element or elements as points of prominence within the message.... Information focus is one kind of emphasis, that whereby the speaker marks out a part of a message block as that which he wishes to be interpreted as informative.” The information focus is what he called the “new”. In contrast to “new”, the term “given” refers to the element which is not new. In the context of a discourse, given denotes the elements which has been presupposed in the forgoing discourse.

Halliday (1967) is particularly concerned with the organisation of information within spoken English and the realisation of this organisation by phonological traits. For example, tone groups mark the border of information units and intonations signal the information focus. More recently, many scholars have extended the discussion by studying the relations between sentence syntactic structure and the given-new structure (Brown and Yule 1983, p. 153-175; Prince, 1992; Gundel, 2003). For example, Prince (1992) remarks that subject NPs tend to be

syntactically definite and old and object NPs tend to be indefinite and new. Gundel (2003) provides a givenness hierarchy, as shown in the below chart. The cognitive statuses on the givenness hierarchy (in the first line of the chart) represent referential givenness statuses that an entity mentioned in a sentence may have in the mind of the addressee.

In	>	activated	>	familia	>	uniquely	>	referential	>	type
focus				identifiable						
<i>it</i>		<i>this/that/this N</i>		<i>that N</i>		<i>the N</i>		indefinite		<i>this N a N</i>

The given-new system differs from the theme-rheme system in that it is about the structure of information unit instead of a clause. There are strong correlations between given-new and Halliday’s theme-rheme system. According to Halliday (1967), in an unmarked sentence, the theme is given and the rheme is new. While theme takes the initial position of a sentence, the given-new system also has something to do with order. For example, according to Hutchins (1977a), in a typical sentence, elements of the theme (similar to Halliday’s given) will precede elements of the rheme (similar to Halliday’s new). In general, “communicative important (contain more new value) will tends to order behind communicative less important element “. (Hutchins 1977a, p. 20)

Some scholars also use term ‘topic’ to refer to what given represents here. For example, Vallduví 1990, p. 44) studies an earlier work by Hajičová and Panevová (1986) and proposes a topic-focus distinction. “The definition of topic and focus are phrased in terms of contextual boundedness or contextual freeness” (Vallduví 1990, p. 44). Van Dijk (1977, p. 117) uses topic-comment similar to the given-new distinction. He remarks that “it may be said that topics are those elements of a sentence which are BOUND by previous text or context. ... The topic-comment structure essentially is a structure relating to the referents of phrases: in general a phrase is assigned topic function if its value in some possible world has already been identified as a value of expressions in preceding implicit or explicit con-textual propositions.”

2.2.1.4 EXTENSIONS OF TOPIC-COMMENT

Here we use topic-comment to generally refer to the above-mentioned three systems.

Traditional views of the topic-comment system define one information structure partition within a system, i.e., a sentence is composed of one theme and one rheme. Recently some researchers propose multiple information structure partitions within a sentence, such as in Kruijff-Korbayová and Webber (2001). Komagata (2001) studies Partee’s (1996) work and

proposes an embedded information structure. Here we will not discuss these theories in details.

Up to now, we make a clear-cut distinction between topic and comment in a sentence. Givon (1983, 1988) argues that “unitary, functional notion” of topic, which has been accepted since the 1970s, must be discarded. According to him, it seems clear that there are *degrees of topicality* marked by different syntactic structures. In addition, within the same syntactic structure, the degree of topicality varies with the NP position: for example, it has been observed that a direct object was somehow “more topical” than an indirect object. Givon (1988) further captures measurement of topicality in terms of continuity of reference, predictability and accessibility. Detailed discussion is out of the scope of this thesis.

2.2.2 SUMMARY OF SENTENCE TOPIC

In the above-mentioned three systems, the notion of topic is proposed to define the structure of a clause or an information unit. Here we will not attempt to clearly define the differences among the three systems. We will focus on the role of topic in information (knowledge) indexing and sharing. This focus aligns with the general aim of the thesis, which is to understand the structure of user queries and to facilitate finding relevant information to user queries. Below are a few major points drawn from the theories of sentence topic that are relevant to our focus.

First, the topic of a clause typically refers to an *entity* in a knowledge representation system. As noted in the above theories, the topic of a clause is typically the grammatical subject and therefore is a noun phrase, while the comment is a predicate. A noun phrase usually refers to an entity in a knowledge representation system and a predicate refers to an assertion made about the entity, which either sets the value of a particular *entity attribute* or linking the entity with another one by a particular *relationship*. It is true that a noun phrase could also refer to some abstract concepts such as ‘the social status of women in Arab countries’. Here in this thesis we will simplify the discussion by focussing on the easy cases where topic refers to an entity.

Secondly, when sharing a piece of new information, topic plays the role of connecting to the recipient’s current knowledge space. The given-new system states that it is a typical structure for an information unit to contain a part that is *known* to the hearer and a part that is *unknown*. Also as mentioned in the discussion of the theme-rheme system, when selecting the theme, the speaker has to take the hearer’s state of knowledge (what s/he knows) and the

hearer's expectation (what s/he understands is being talked about) into account. For example, the sentence 'Mary, John kissed her' is an appropriate answer to question 'what happened to Mary'; in contrast, sentence 'John has kissed is Mary' seems to be appropriate to answer question 'who has John kissed'. Thus, we can infer that an effective way of sharing new information is to start with something that already exists in the current knowledge space of the recipient, which is the role that topic plays. One could argue that there are cases that the speaker talks about something that by itself is new to the hearer. For example, the speaker talks about a person called 'Napoleon' that the hearer has never heard about. However, in such cases, the speaker may start with 'in French history, there is a great man called Napoleon'; by saying so, s/he presumes that the hearer knows about France and the concept history. To summarise, in the process of knowledge sharing, topic plays the role of connecting the message that the writer/speaker wants to deliver to the recipient's current state of knowledge. It is worth noting that in an information system that aims to retrieve relevant information for different user queries, the same piece of information may have different topics to different users because their knowledge spaces are different, which may also be different from what the author originally intends to be the topic.

2.2.3 THEMATIC ANALYSIS

Thematic analysis refers to automatically identifying the theme-rheme structure for a clause or a sentence. The notion theme here refers to either Halliday's (1967) theme or Firbas' (1964) theme. Identifying the thematic structure of a sentence may help many applications. According to Komagata (1999), thematic analysis "*has been recognized as a critical element in a number of computer applications: e.g., selection of contextually appropriate forms in machine translation and speech generation, and analysis of text readability in computer-assisted writing systems*". Since thematic analysis does not introduce new definition of topic, therefore we would just briefly introduce two work.

2.2.3.1 DUCHASTEL ET AL. (1992)

Duchastel et al. (1992) do thematic analysis on the Quebec Budget speeches from 1934 to 1960, which is in French and contains various forms of political discourse originating in different Quebec institutions. They aim to show that the pattern of applying different sentence thematic structures is reflective of the social transformation taking place.

Following Halliday's (1967) definition of the notion of theme, they take the semantic elements in the first position of the clause as the theme. They use the GDSF parser⁴ to parse sentences in the corpora and extract words of the themes. They compare the ratios of words in themes among discourses of different sociological families, such as, Economics, Politics and Social Institutions. Further, themes are categorised into two sets, marked themes and unmarked themes. Their analysis shows that there is an apparent rise of the ratio of marked theme during the 1945 to 1948 period in General Economy, Natural Resources and Industries. They attribute the rise to the political movement during that period of time.

2.2.3.2 KOMAGATA (1999)

Komagata (1999) presents a system which translates English to Japanese. The aim of his thesis is to show that knowing the thematic structure of English will help choosing appropriate particles in Japanese.

His thesis “adopts a classic theory of information structure as binomial partition between theme and rheme, and captures the property of theme as a requirement of the contextual-link status.” “The notion of ‘contextual link’ is further specified in terms of discourse status, domain-specific knowledge, and linguistic marking. ... The identification process can then be specified as analysis of contextual link status along the linguistic structure”. They use Combinator Categorical Grammar⁵ to get surface syntactic structure.

The machine prediction of particles is compared against human translations. The evaluation results demonstrate that the prediction results are much better than other approaches, including baseline systems such as random selection, approaches that only utilise contextual status and those that only utilise structural information and linguistic markings.

2.3 TOPIC AT THE DISCOURSE LEVEL

Above we introduced that the notion of sentence topic provides an account of sentence structure or information structure. At the discourse level, the notion of topic is often discussed in the context of investigating issues such as how a discourse is organised, why it appears coherent and how to generate a concise representation of the discourse content. A widely adopted term in the literature is *discourse topic* or *conversational topic*. This section is divided into two parts: section 2.3.1 talks about some formal theoretical work on discourse

⁴ Refer to Plante (1980)

⁵ Refer to Ades and Steedman (1982)

topic; section 2.3.2 introduces a few work that aim at automatically generating topic expressions to solve practical problems.

2.3.1 THEORETICAL WORK ON DISCOURSE TOPIC

Below we will introduce several theories, including Brown and Yule's (1983) *topic entity*, van Dijk's (1977) *discourse topic*, Hutchins' (1977a) *aboutness* and Brown and Yule's (1983) *topic framework*.

2.3.1.1 BROWN AND YULE'S (1983) TOPIC ENTITY

A simple definition of discourse topic is the most frequent sentence topics in the discourse.

Brown and Yule (1983, p.135) discuss the earlier work of Katz (1980) and define that "the notion of a discourse topic is that of the common theme of the previous sentences in the discourse, the topic carried from sentence to sentence as the subject of their predication". Here theme loosely refers to the grammatical subject of a sentence. Brown and Yule (1983) comment that one basic organisational method for discourse production involves placing the main referent in the subject position. This introduces a formal process called thematisation. They discuss the earlier work of Perfetti and Goldman (1974) and define thematisation as "the discourse process by which a referent comes to be developed as the central subject of the discourse." They use "topic entity" to refer to such a thematised referent.

De Beaugrande's (1980) approach resembles the above thinking. He proposes to use a *conceptual network* to represent textual content. He defines a list of conceptual relations (e.g., state-of, substance-of and reason-of), which link between elements of a sentence to construct a network. By representing the textual content in such a way, he finds that one node in the network is shared by all the sentences. He calls such a node as the "Topic". Different from the above definition, he does not mention whether the most frequent element is also the most frequent sentence topic.

2.3.1.2 VAN DIJK'S (1977) DISCOURSE TOPIC

Van Dijk (1977, p. 49 – p. 136) proposes a notion of discourse topic in the context of discussing how a discourse is organised. He notes that facts denoted by sentences are related with respect to some "COMMON BASIS" or from a certain "POINT OF VIEW" (van Dijk 1977, p. 49). The "COMMON BASIS" or the "POINT OF VIEW" is what he calls the discourse topic. He further defines discourse topic as a proposition which is *entailed* jointly by the discourse sequence within the text segment. A sample text he gives talks about the economic status of a small

town named Fairview. It includes some details about the factory, the quality of the food and the competition from another town. The details jointly imply that the town is enduring an economic hardship which was not the case before. The topic of the sample text is defined as “a (little) town (called Fairview) is declining because it cannot compete with another town (called Bentonville)” (van Dijk 1977, p. 134). The concept ‘decline’ is inferred because the details in the text matches the *frame* associated with this concept, i.e., a subsystem of knowledge which includes typical phenomena of economic prosperity and non-prosperity. He further states that expressions such as ‘a town’, ‘two towns’, ‘the competition between two towns’ are not as qualified as the above-proposed topic since a discourse topic must “DOMINATE all semantic information of the sequence” (van Dijk 1977, p. 136). Detailed discussion of the definition of the notion *dominate* is out of the range of this thesis⁶. We think that the above definition for discourse topic is too strict; especially it is not appropriate to require the detailed facts in the discourse to entail the discourse topic since entailment is a formal logic relation. Instead of using the term entailment, we might only be able to say that the kind of phenomena⁷ described in the sample text is representative for a declining town, hence the concept ‘decline’ would be an enough precise characterisation of those details.

Based on this notion of discourse topic, van Dijk (1977) further presents the macrostructure of a discourse featured by a hierarchical structure with each node denoting a discourse topic which organises a text segment. Thus, the higher a discourse topic is in the hierarchy, the larger the textual segment that it organises. He also defines an operational model — i.e., the macrorules — in the process of discourse topic induction, which mainly include select, reduce and generalisation operations. Above we have used the word ‘decline’ to illustrate the generalisation operation. The select and the reduce operation, according to Hutchins (1977a, p. 30), may be defined as the elimination from a semantic network of those elements and relationships which are inessential and unimportant to the development of the main plot or argument. In general the frequently occurring participants and activities are more likely to be kept in the macro-structure. We may notice that the strategy of keeping the most important and frequent element is quite similar to the topic entity approach.

⁶ We actually do not think the logic expression used by van Dijk (1977: 136) for defining the notion DOMINATE is valid

⁷ Phenomena like shabby house, unkept road, low quality food, etc.

2.3.1.3 HUTCHINS' (1977A) ABOUTNESS

Hutchins' (1977a) study falls into the context of an information system. In particular, he explores how to pick up elements in a discourse to make good indices.

He studies text structure and identifies two basic types of sentence progression inside a discourse from the viewpoint of theme-rheme articulation: "linear progression, where the favoured thematic elements relate to elements of a preceding rheme; and parallel progression, where theme remains constant". He applies the theme-rheme distinction at the sentence level to the discourse level. Specifically, he asserts that in each type of sentence progression, "the first sentence provides the starting point or foundation for the following sentences. In this sense it may be regarded as a whole as the theme of the paragraph, where the subsequent sentences contribute the rheme."

Hutchins (1977a) further defines a notion of "aboutness" to address the problem of making good indices of a discourse. He notes that "Whenever anyone consults an information system in search of a document answering a particular information need, he cannot in the nature of things formulate with any precision what the content of that document should be". He argues that the aboutness should be defined as the thematic element of a text, rather than being representative of any new information. Therefore, "the 'aboutness' of a document is to be sought in those initial sections" where the author "establishes points of contact with what he assumes to be the 'states of knowledge' of potential readers". We see that Hutchins' aboutness is essentially different from van Dijk's (1977) discourse topic, with the discourse topic being defined as a proposition dominating all the discourse details and the aboutness being defined as the thematic elements only.

2.3.1.4 BROWN AND YULE'S (1983) TOPIC FRAMEWORK

Brown and Yule (1983, p. 68 - p. 121) criticise many former theories which define the discourse topic to be one single correct expression. They argue that "for any practical purpose, there is no such thing as the one correct expression of the topic for any fragment of discourse. ... What is required is a characterisation of 'topic' which would allow each of the possible expressions, including titles, to be considered (partially) correct, thus incorporating all reasonable judgement of 'what is being talked about'". Instead of using one possible phrase or clause, they argue that we should have a bunch of features extracted from the context and the speaker and hearer's shared knowledge and the ongoing discourse to characterise the current situation, which is the topic framework. Certainly those features should not be any

shared knowledge between the hearer and the speaker; instead they must be some activated features. To judge whether a piece of detailed information is relevant to the topic framework, they note that “a discourse participant is speaking topically when he makes his contribution fit closely to the most recent elements incorporated in the topic framework” (Brown and Yule, 1983, p. 79). Here we want to point out that their study focuses on casual conversations, the topic of which might be very different from that of a formal written work. Specifically, in a casual conversation, people could talk without a defined aim; besides that, the speaker and the hearer could exchange their roles and lead the topic into different directions.

2.3.2 AUTOMATIC TOPIC ANALYSIS AT THE DISCOURSE LEVEL

Topic is also broadly studied in the research of computational linguistics, where different research areas, such as information retrieval (IR), topic detection and tracking (TDT), text classification (TC), text segmentation, topic generation and text summarisation, may have different interpretations of the notion. However, studies in these different areas resemble each other in the sense that they all need to capture and encode the topic in a certain structure as a characterisation of the discourse content.

An IR system helps to retrieve information from a large document collection relevant to user queries; a TDT system or a TC system automatically classifies documents into different predefined topics or categories; a text segmentation system segments a text into several parts, each of which is on a slightly different subtopic. There is no requirement for generating explicit topic expressions in these systems. To capture the topic of a discourse, these systems usually do not use complex linguistic analysis; instead they simplify the view of the textual content as a list of key words weighted based on word distributional characteristics. In comparison, automatic text summarisation or topic generation systems aim to generate a more concise representation of the original discourse content. Typical approaches fall into two categories: those based on template instantiation (Lin and Hovy, 1998) and those based on passage extraction (Boguraev and Kennedy, 1997; Kan et al., 2001). The template-instantiation-based approach typically contains several steps, including preparing a set of predefined templates, analysing the discourse content to identify which template it fits in, extracting elements from the discourse to fill in the information slots defined in the templates and generating summaries. Instead of using this rather complicated approach, most systems (Boguraev and Kennedy, 1997; Kan et al., 2001; Arora and Ravindran, 2008) take the approach of directly extracting the most important and representative element, by means of a notion of

salience. Salience is calculated based on shallow features such as sentence positions, term frequencies, cue phrases, strength of lexical chains, etc. In recent years, the research on automatic text summarisation is mainly driven by the DUC⁸ and the TAC⁹ conferences. One of the tasks defined by the DUC and the TAC conferences is to generate multi-document summaries for news articles. Majority of the participants generate summaries by directly extracting sentences from articles. Similar to IR, the relevance of a sentence is determined by the word distributional information. One prevalent approach is the vector space model approach (Gong and Liu, 2001). The importance of a sentence is calculated based on the inner production between the sentence vector and the document vector.

Practical work in the above-mentioned areas usually does not have direct relationships with the theoretical work. According to van Dijk's (1977) macrorules, the process of generating a discourse topic involve both reducing operations and generalisation operations. We would argue that the salience element detection and discourse pruning technique used in text summarisation might embody the idea of reduction. However, the current techniques are still not capable of effectively simulating the generalisation operation that human beings are able to. By saying so, we do not mean that the current approaches could never generate a good topic; often good topic expressions are already in the text and could be identified using simple cues. Below we will introduce a few work in text summarisation and topic generation.

2.3.2.1 NORRIS' (1998) ESSENCE

Norris (1998) differentiates between two modes of re-presentation of textual information, i.e., gist and aboutness.

"The gist of a text refers both to what a text is about, and to the arguments contained within any exposition, and any conclusions that may be drawn from these arguments. It is restatement of the main points and the 'thrust' of the text, and gives an idea of the flow of information through the text. ... On the other hand, the aboutness of a text, or 'what the text is about' serves to indicate the content of the text. ... The aboutness need not tell us any information about how information flows, or is developed, throughout the course of the text"
(Norris 1998, p. 22)

⁸ Refer to <http://duc.nist.gov/>

⁹ Refer to <http://www.nist.gov/tac/>

She further asserts that aboutness is normally represented by key expressions (e.g, key named referents such as people and places) or indexing terms, which are selected from the text by virtue of their importance. She points out that key expressions and indexing terms are often too general to represent the text. Instead, she proposes to generate a kind of compound nominals which are equally concise but are more informative. The detailed approach contain several steps: first extract all the compound nominals in the original text; among all the heads of the extracted compound nominals, a most *salient* head is selected based on frequencies and how well it connects to other heads; finally, output the salient head together with both its own modifiers and the modifiers of the heads that has strong connection with it. Here connection between two heads is calculated as the number of words that occur in both sentences containing the two heads respectively. Note that the output of the system is not exact compound nominals.

2.3.2.2 BOGURAEV AND KENNEDY'S (1997) TOPIC STAMP

Boguraev and Kennedy (1997) present a domain- and genre-independent approach for text content characterisation. Their system automatically generates a “capsule overview” for a discourse. The capsule overview is composed of a list of linguistic expressions – i.e., the topic stamps – and a specification of the relational contexts (e.g. verb phrases, minimal clauses) in which these expressions appear. Topic stamp is a phrasal expression extracted from the original discourse based on the salience measurement. The concrete method employed embodies the idea that the frequency of a linguistic element is essential to how salient it is. A syntactic and grammatical position based anaphora resolution approach is used to build the co-reference chains. This improves the accuracy of the frequency calculation. The authors use a Forbes article¹⁰ as an example and summarise topic shifts of this article as “relation between Apple and Microsoft”, “future of desktop computing” and “Amelio’s background and plans for Apple’s operating system”; in comparison, the corresponding topic stamps generated by the system are “apple: Microsoft”, “desktop machines: operating system” and “Gilbert Amelio: new operating system” respectively. Comparing the results generated by the system and the author’s own topic expressions, it is obvious that this approach cannot simulate the process of abstracting abstract concepts such as “future”, “plan”, “background” and “relation”.

¹⁰ Part of this article is presented as example [1.5] in chapter 1.

2.3.2.3 RINO AND SCOTT'S (1996) GIST

Rino and Scott (1996) propose a discourse model for automatic text summarisation. This model contains three essential elements: the communication goal, the central proposition and a knowledge base. The knowledge base is a hierarchical structure in which the leaf nodes encode the propositions in the discourse and the intermediate nodes encode the semantic relationships. The summarising process consists of two pipeline steps: prune the original hierarchical structure and reorganise the remaining information into coherent textual expressions. The process ensures that the original communication goal is fulfilled and the central proposition is kept in a prominent position, which suggests that the communication goal and the central proposition are the *gist* of the discourse. Discourse planning operators are defined to fulfil these two steps. These operators map high level communication goals (e.g., describe, discuss and expose) or linguistic goals (i.e., primitive acts such as inform or rhetorical goals such as evidence and justify) to lower level ones. In the operator definition, the low level goals are not equally important. The importance of a goal is the criterion for determining whether the goal should be implemented in the summary. Intentional and semantic constraints are defined as the preconditions of the application of a planning operator.

We would say that this approach is appropriate for the kind of discourse which contains clear “logic chains” (Rino and Scott, 1996). Such kind of discourses often contain a central proposition and has clear logical inference in the flow of the text.

2.3.2.4 MARCU'S (1997) DISCOURSE STRUCTURE APPROACH

The rhetorical structure theory (RST) (Mann and Thompson, 1987) defines the organisation of a discourse as a tree structure. The leaf nodes of the tree are clauses, which are linked at upper levels by rhetorical relationships, such as motivation and evidence. Many rhetorical relationships are asymmetrical: of two text spans that hold a RST relationship, one is more important than the other. For example, the evidence relation links two text spans, with one providing evidence for the statement included in another; the latter is more important than the former one. A notion called nuclearity is used to refer to this attribute. Marcu (1997) aims to generate discourse summaries by keeping the most important text spans based on the RST theory. He did an experiment on articles from Scientific American¹¹ to verify the idea that

¹¹ Refer to <http://www.sciam.com/>

salient elements of a discourse can be extracted based on the notion of nuclearity. Human subjects are hired to rank sentences according to their importance in representing the discourse content. The ranking result tends to agree with the importance of a sentence indicated by its position in the rhetorical structure. The experiment result supports the above hypothesis. This approach, similar to the approach used in Rino and Scott (1996), applies better to evaluative texts than to narrative or descriptive text. We could well imagine that the sequence relation and the joint relation are two predominant relationships in narratives. These two relationships are multi-nuclear and two text spans linked by multi-nuclear relationships are equally important.

2.3.2.5 SCRIPT BASED APPROACH

The notion of *script* is proposed by Schank and Abelson (1977). According to Mani (2001, p. 34), a script is “a specialized schema which identifies common, stereotypical situations in a domain of interest”. An example could be a scenario of eating in a restaurant which usually subsumes a sequence of actions such as sitting down, reading the menu, ordering, eating, paying, and leaving. It has been shown by Mani 2001 (p. 34, 130) that when newspaper editors summarise a newspaper article, they try to identify a high-level schema, such as “supermarket shopping”, “travelling by train”, etc. The identification of such high-level scripts is implemented in DeJong’s (1982) FRUMP and later in Lin and Hovy (1998). In Lin and Hovy (1998), scripts are triggered by the identification of key words. For example, words like “table, menu, waiter, order, eat, pay, tip” are good hints for the script “restaurant-visit”. In FRUMP (DeJong, 1982), some other inductive rules are applied to trigger a script. For example, “world knowledge that arrests usually follow crimes would be used to trigger the arrest script once a crime script has been triggered” (Mani 2001, p. 131). Lin and Hovy (1998) combine the script-based approach with some typical text summarisation techniques. Specifically, they also select salient elements from the discourse based on term frequencies, sentence positions and cue phrases. We see that the operations used there are quite like the macrorules defined in van Dijk (1977). In particular, the step of selecting elements corresponds to the selecting and reducing operation; the step of abstracting scripts corresponds to the generalisation operation.

2.3.2.6 LATENT TOPIC MODEL APPROACH

Practical research areas such as Information retrieval and text summarisation typically simplify document representation as a set of key words associated with word distributional

information. One model is the vector space model, which represents a document as a vector of real numbers, each dimension corresponds to one key word and the value of which is calculated based on word occurrence in a document. The details of the vector space model and its application in IR will be discussed in section 6.2.3. Gong and Liu (2001) first apply the vector space model in text summarisation. They generate summaries for a document by directly extracting sentences from the document. To find which sentence best represents the document content, they compute the inner product between the sentence vector and the document vector. One key problem with the vector space model is the sparseness of the vectors; another problem is that the correlations among words are not captured with each word being represented as one vector. Gong and Liu (2001) apply latent semantic analysis to reduce the dimensions of the vectors. The reduced dimensions are orthogonal to each other and are supposed to capture the latent semantics in the document.

Another document representation model is the Latent Dirichlet Allocation (LDA) model proposed by Blei et al. (2003). LDA is a generative probabilistic model of a corpus. The basic idea is that documents are represented as random mixtures over latent topics, where each topic is characterised by a distribution over words. They model the probability for a latent topic to generate a word and the probability for a document to contain different topics. Further they can derive a document representation based on the probability for a document to generate a word. Chang and Chien (2009) apply this representation to generate document summaries. They extend this model to sentence level and derive a sentence representation as the probability for a sentence to generate a word. They calculate the Kullback-Leibler (KL) divergence between the sentence representation and the document representation to select sentences that best represent the content of the document.

2.3.3 SUMMARY OF DISCOURSE TOPIC AND A TOPIC CLASSIFICATION FRAMEWORK

At the discourse level, the notion of topic is defined as a concise representation of the discourse content or as a discourse organisational principle. These two roles are closely related, for that topic expressions representative of discourse content must also have the function of organising the detailed elements in the discourse and vice versa. As mentioned in section 2.2.2, we focus on analysing the function of topic in the process of information (knowledge) indexing and sharing; thus, the first sense is more relevant to our point of view.

Most previous approaches in finding the topic of a discourse, as introduced in section 2.3.1 and section 2.3.2, reflect a structural view of the discourse content. Specifically, a discourse contains some topical elements and the comments about them. Many theories believe that it is a typical structure of a discourse to centre around one entity or a few entities, which is the topic of the discourse; this view is reflected in Brown and Yule's (1983) topic entity, in Boguraev and Kennedy's (1997) topic stamp and in Norris' (1998) aboutness. As previously noted, to find the topic of a discourse, most practical systems in the area of IR, topic generation and text summarisation take word frequencies as an important indicator of how representative the word is to the content of the discourse. A slightly more sophisticated approach is to also take syntactic positions into account. For example, Boguraev and Kennedy (1997) give more weight to phrases at the subject position; Brown and Yule (1983) also note that it is a typical discourse organisational principle to put the main referent in the subject position. In addition to the above-mentioned discourse structure, in which most sentences in the discourse sharing the same theme, Hutchins (1977a) proposes another typical discourse structure called "linear progression", where the theme of each sentence relates to the rheme of the preceding sentence. Similar to us, he aims to facilitate locating relevant information by generating effective indices to a discourse. For this purpose, he argues to index a discourse using the sentence topics in the initial sections so as to establish points of contact with the 'state of knowledge' of potential readers. Here we see that similar to sentence topic, topic at the discourse level is also defined by some discourse theories as having the function of connecting to the reader's knowledge space.

In many of the above-mentioned theories, a discourse topic not only contains the centred entity, but also contains the key comments made about this entity. For example, in Boguraev and Kennedy (1997), they first identify the "topic stamps" and then generate a "capsule overview" of the discourse content by attaching the verb phrases or clauses where the topic stamps occur; in van Dijk (1977), the discourse topic is a proposition that "dominate" all the information in the discourse, which is generated by selecting the most important elements and relationships from the semantic network representing the content of a discourse.

Based on the above discussion, we would then apply the given-new distinction to the discourse level, that is, a discourse adds new knowledge to some known elements. Further, a distinction could be made between topics that only contain the known elements and those that contain both the known elements and the new information added. In the above-mentioned theories of discourse topic, the centred entity is most likely to be a known element

and the comments about the entity are new. In terms of the function of knowledge indexing, a topic expression that contains the topic elements has the function of indicating the discourse content, while topic that includes both the topic elements and the comments has the function of reproducing the discourse content.

This distinction could be aligned to the distinction between indicative summary/abstract and informative summary/abstract in the area of text summarisation and text abstraction. Norris (1998: p. 28, 29) remarks that “an indicative abstract is written with the aim of indicating what a text is about, rather than actually imparting any of the specific information given in the original” and informative abstract “is aimed at reproducing the gist of the original text”. According to Kan et al. (2001), “an indicative summary's main purpose is to suggest the contents of the article without giving away detail on the article content. It can serve to entice the user into retrieving the full form” and “an informative summary is meant to represent (and often replace) the original document. Therefore it must contain all the pertinent information necessary to convey the core information and omit ancillary information”. We therefore call the first type of topic expressions *indicative topic* and the second type *informative topic*. Below we will try to probe some characteristics of these two types of topic.

From the syntactic point of view, the difference is that an indicative topic is typically a nominal phrase while an informative topic expression is typically a clause. A nominal expression is the linguistic representation of entities or things about which a text is going to say something (Norris, 1998); a clause usually sounds like making a statement and therefore informs something new. Below is a pair of indicative abstract and informative abstracted given by Norris (1998: p. 29). In the indicative abstract, the actual topic expression (marked in grey) is embedded into some typical phrasal frames such as ‘... is discussed’ and ‘... are mentioned’. It could be observed from this example that indicative topic expressions are usually nominal phrases while informative topic expressions are all clauses.

[2.5] a. An informative abstract

The video games industry is growing fast and will dominate the toy market and become an established part of home entertainment. The 1991 computer games market was worth 275 million pounds sterling growing to 500 million in 1992, half the toy market. Hardware sales will rise from 261 to 635 million pounds sterling in 1994. Associated software sales are forecast at 645 million pounds sterling in 1993. The compact disc market is worth 345 million pounds sterling. The main

competitors in the market are Sega and Nintendo. Nintendo will spend 15 million sterling on advertising over Oct-Dec 1992.

b. An indicative abstract

The growth of the video games market is discussed, with particular reference to the actual and expected increase in its value over the period from 1991-1994. The main competitors are mentioned, with an example of the advertising budget of one of them.

It is worth noting that we could always transform a clause into a nominal phrase by normalising the predicate. For example, the clause 'Fairview is declining' could be transformed to 'the decline of Fairview'. Thus, it seems that the syntactic difference does not make any difference in the amount of information that is conveyed. Nonetheless, we do think that different syntactic forms make the topic expression sound different. An event or a fact is usually presented in the form of a nominal phrase if it is something known; instead, when it is something unknown, it is usually be expressed in a statement.

What interests us is indicative topic because it contains known information only and therefore could well connect to the reader's state of knowledge. Under the context of an information system, the system could index a discourse by the indicative topics of it; at the document retrieval stage, it could match the indicative topics to user queries to retrieve relevant documents.

As mentioned above, most theories and practical systems approach topic elements as the most frequent term or the most frequent sentence topic. However, we see in the above examples, the indicative abstract contains not only the topic elements of the informative abstract but also a general characterisation of the comment. As a simplification, we mark the subject of each sentence in the informative abstract in bold face as the topic element and mark the predicate in grey as the comment. We see that there are direct references to the topic elements in the indicative abstract, such as 'video games market' and 'main competitors'. We also see terms in the indicative abstract that relate to the comments in the informative abstract. For example, 'value' relates to concrete numbers such as '275', '500' and '261'. We could relate this observation on topic to those given in chapter one, where we show that many topic expressions contain a part abstracted from the detailed content in the discourse. We will further investigate on these observations and will develop a theory of extended topic in the next chapter to achieve a better formulation of indicative topic structure.

[2.6] a. An informative abstract

The video games industry is growing fast and will dominate the toy market and become an established part of home entertainment. **The 1991 computer games market** was worth 275 million pounds sterling growing to 500 million in 1992, half the toy market. **Hardware sales** will rise from 261 to 635 million pounds sterling in 1994. **Associated software sales** are forecast at 645 million pounds sterling in 1993. **The compact disc market** is worth 345 million pounds sterling. **The main competitors** in the market are Sega and Nintendo. **Nintendo** will spend 15 million sterling on advertising over Oct-Dec 1992.

b. An indicative abstract

The growth of the **video games market** is discussed, with particular reference to the actual and expected increase in its value over the period from 1991-1994. **The main competitors** are mentioned, with an example of the advertising budget of **one of them**.¹²

2.4 SUMMARY

In this chapter, we provide a broad review of the work on topic. The discussion is divided into two levels, i.e., topic at the sentence level and topic at the discourse level.

At the sentence level, topic is defined as an important constituent in sentence structure analysis. A distinction is made between topic and comment, with topic typically corresponding to the grammatical subject of the sentence and comment is the predicate. There are two closely related systems, the theme-rheme system and the given-new system. We compare the differences among the three systems. As mentioned at the beginning of this chapter, our focus is on the role of topic in knowledge representation and knowledge sharing. We derive from the existing theories that sentence topic is typically an entity in a knowledge representation system. In the process of knowledge sharing, we infer that an effective way for knowledge sharing is to have one part connecting to the recipient's knowledge space, which is the role that topic plays.

¹² The numbers in example [2.6] are used to mark the mapping relations between elements in the informative abstract and in the indicative abstract.

Chapter 2 – A Review of Theories of Topic and a Topic Classification Framework

At the discourse level, the notion of topic is developed under the discussion of how a discourse is organised or how to generate a concise representation of discourse content. Some theories generate a discourse topic by extracting the most frequent sentence topics (or the most frequent terms as a simplification). Some other theories further attach important comments either inferred or extracted directly from the discourse. We use indicativeness and informativeness to mark this distinction. We are more interested in indicative topic because it only contains known information and therefore could directly match to user queries to find relevant information. We compare the topic expressions contained in an indicative summary and in an informative summary. The comparison shows that indicative topic not only contains the topic element of the corresponding informative topic but also contains a general characterisation of the comment. None of the theories introduced in this chapter explicitly addresses this observation. We will further explore the structure of indicative topic in the next chapter.

CHAPTER 3

THEORY OF EXTENDED TOPIC

3.1 INTRODUCTION

In last chapter we introduced many different theories which provide different interpretations of the notion of topic: as an important sentential element in sentence structure analysis; as a concise re-presentation of the discourse content; and as a discourse organisation principle. The second sense is more related to the practical aim of the thesis, which is to improve finding relevant information to user queries. In section 2.3.3, we introduced the distinction between indicative topic and informative topic and further focused our discussion on indicative topic. We noted that most existing theories of topic as well as practical topic generation systems define discourse topic as the focused entity (i.e., the most frequent and prominent entity or sentential topic). In contrast, we showed that human-generated topics also contain elements characterising the comment on the focused entity. Below we will use another example (example [3.1]) cited from (Parouty, 1993) to illustrate the linguistic phenomenon under discussion.

[3.1] Passage 1

Mozart was born in January 27, 1756 in Salzburg, Austria. At three years old Mozart began to play the harpsichord; by six he was writing compositions. His father Leopold decided not to waste Wolfgang's precocious talents and took him on a tour across Europe with his sister, displaying their abilities to royal courts and houses. These tours continued well into Mozart's late teenage years.

Passage 2

After flirting with a first cousin and being attracted to Aloysia Webber, Mozart opted for Aloysia's sister, Constanza. Temperamentally they were perfect, both fun-loving and playful; financially, however, they (especially Mozart) were spendthrifts. After a long wait, with Leopold's blessing not forthcoming, they decided to marry on August 4, 1782. Their marriage had ups and downs, but lasted until Wolfgang's death in 1791.

Following the above-mentioned topic generation approach, we might say that both of the passages above are about Mozart (or perhaps Mozart and Constanza for the second passage). However, the simple (and true) observation that both passages are 'about Mozart' obviously leaves out something important: the perspective from which the focused entity is approached. We want to say that passage 1 is about Mozart's childhood, while passage 2 is about his marriage. One could imagine further passages about his musical compositions, his employment, his significance today, and so forth. The labels 'childhood' and 'marriage' provide a strong indication of which fact belongs in which passage – probably sufficient to allow exact agreement between two judges.

The aim of this chapter is to achieve an improved formulation of topic by developing a theory of extended topic. By 'extended topic' we mean a topic definition comprising a perspective as well as a focused entity. We will provide a formal definition of perspective in section 3.2; section 3.3 provides a list of contexts in which extended topic occurs; section 3.4 summarises the intuitive basis of the idea and why it might have practical value.

3.2 THEORY OF EXTENDED TOPIC

In previous discussions we have defined our focus on indicative topic. The function of indicative topic is to help set up the right expectation of the discourse content. Specifically, by reading an indicative topic, people would understand the range of information contained in the discourse and would be able to judge whether it is what s/he is looking for. In this sense, indicative topic is similar to user-formulated questions, which have the function of defining what facts are relevant.

The notion of extended topic could be formalised by probing the characteristics of user queries to a knowledge base. It is important to change our focus from indicative topic to user queries. First, the change of focus may simplify the problem since queries have a clear function. We will talk about it in more details later. Secondly, understanding user queries has more practical value. In particular, analysing the connection between user queries and the detailed elements in a relevant discourse may help improve automatic document retrieval and topic generation.

It is clear from the above change of perspective that a topic formulation is not a statement: it is not designed to impart factual information. Two qualifications are necessary here. First, a topic formulation like 'marriage of Mozart' might inform some readers that there existed a man named Mozart and that he was married. However, the normal expectation would be that

these facts are presupposed (i.e., already known). Second, there is one sense in which a topic formulation is informative: it provides information about a discourse. Paradoxically, not only is the discourse about the topic, but the topic is in a sense about the discourse. A title, or an entry in an index, tells you something about the book you are reading rather than something about Mozart.

More importantly, this topic formulation does not serve to state facts, but serves the function of constraining which facts are selected. When querying 'marriage of Mozart', we expect to find details of Mozart's wife, their children, their domestic life, etc., as opposed to an account of Mozart's childhood, or (to stray much further afield) a recipe for making an omelette. Without knowing most of these details in advance, we can still recognise their relevance. This suggests a way of defining extended topic more precisely: we can think of it as something which operates on a knowledge base in order to select a subset of relevant facts.

To formalise this idea, we might think of a knowledge base K as a set of facts, and a topic T as a function $T(K)$ which yields a subset of K. This formalisation simplifies by classifying each fact as either relevant or irrelevant, rather than allowing degrees of relevance; an alternative would be to think of T as assigning a value between 0 and 1 to each fact, where 1 means maximally relevant and 0 means maximally irrelevant. Which facts were selected would then depend on the desired length of the discourse.

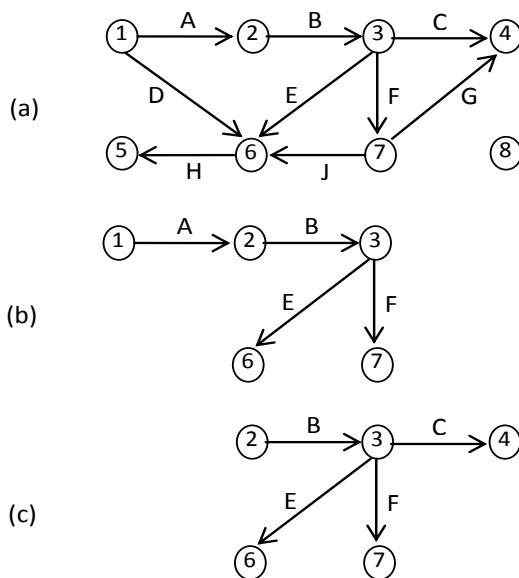


FIGURE 3-1. SELECTING FROM A KNOWLEDGE BASE

Figure 3-1 illustrates the simpler model of topic as selecting a relevant fragment from a knowledge base. Network (a) shows schematically the complete knowledge base, comprising

a set of entities (nodes labelled with integers) and relationships (arcs labelled with letters). The diagram encodes a set of facts $\{A(1, 2), B(2, 3), C(3, 4), D(1, 6), \dots\}$ each of which is a relationship between two entities. Network (b) shows a fragment of (a) selecting four related facts. Following this model, a topic formulation would have to provide criteria allowing (b) to be selected from (a). How might this be done?

Part of the answer obviously lies in the identification of a focused entity (or focused entities). In terms of the diagram, this would simply be one of the nodes — for instance, node 3. If the topic was formulated only through a focused entity, selection would presumably have to depend on whether a fact included this entity; the resulting network is shown in figure 3-1(c). However, it is surely rare to find a discourse in which every fact concerns the same entity, and all known facts concerning this entity are reported. Network (b) is a better schematic model for a discourse about node 3, since the dominance of the focused entity is less complete: $A(1, 2)$ is included, even though it does not mention node 3, while $C(3, 4)$ is omitted. Therefore, there must be another part of a topic expression, which provides a criterion for selecting a region of the knowledge network around the focused entity. This region will always include the focused entity, but may also include facts like $A(1, 2)$ in which the focused entity does not participate directly.

One way (method 1) is to point to a relationship and ask for the thing that holds this particular relationship with the focused entity. Figure 3-2 depicts part of a knowledgebase. One question that could be formed for this knowledgebase is ‘what does cow eats’; this way we select grass using a relationship – i.e. eats – around the focused entity – i.e., cow. Similarly, we might ask ‘what eats grass’, then cow, sheep, grasshopper and many others will be selected.

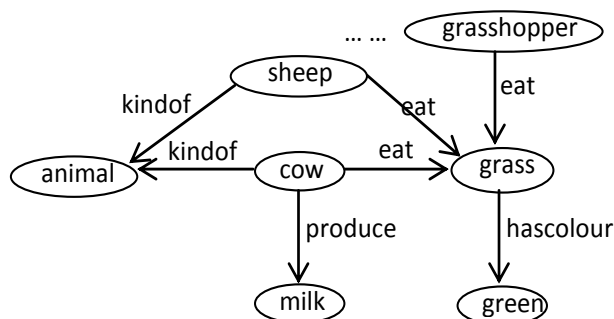


FIGURE 3-2. AN EXAMPLE KNOWLEDGE BASE

It is still rare to ask for everything that eats grass, instead we can further use a general category to confine the kind of things we want (method 2). For example, we may ask ‘what

mammal eats grasses’; this time, only sheep and cow will be selected since they are both a type of mammal.

The above ways only select entities that hold one particular relationship with the focused entity, which is quite limited. Sometimes we will use a concept which primes a bunch of different relationships (method 3). Back to the Mozart example, we may ask ‘the family background of Mozart’. Figure 3-3 depicts part of the answer to this question. The figure shows a conventional family tree rather than a diagram with nodes and arcs, but the same information could easily be encoded using relationships like name, year-of-birth, spouse, and child.

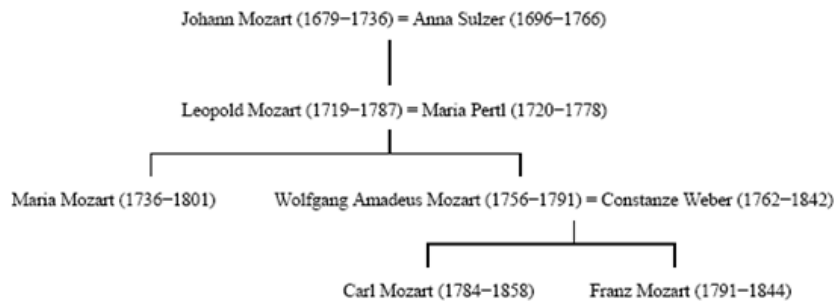


FIGURE 3-3. MOZART FAMILY TREE

Suppose our task is to pick relevant information from a network that contains varied information about Mozart and other composers. Among other things, the knowledge base will mention his main compositions, important events during his career, and the places where he lived. From all this information, mostly irrelevant, we are required to pick a set of facts from which we can put together a discourse on Mozart’s family background. How will we proceed? The obvious first step is that the concept family background will prime a particular set of relationships, those of kinship and marriage shown in figure 3-3. By following relations like child and spouse, we will pick out relevant neighbouring entities such as Mozart’s father, mother, sister, wife, and children. Depending on the amount of detail requested we might go further back, perhaps to Mozart’s grandparents. This gives a skeleton for the discourse, which could minimally present exactly the kind of information found in a family tree.

However, if given this assignment we would probably take a second step, one of elaborating the basic skeleton with a few more facts about each entity. To do this, we might apply to each entity an implicit secondary selecting criteria, that of ‘personal history’. This would prime another set of relationships, including residence, profession, status, as well as name and gender (which are already mentioned on the tree). Thus readers might be interested to know

that Johann Mozart, Wolfgang’s grandfather, was a bookbinder who lived in Augsburg, that father Leopold was an orchestral musician, and that sister Anna was another musical prodigy. The choice of facts here is wider, and a skilled selection would require more sophisticated judgements of what is ‘interesting’. However, the basic procedure is clear and should not be hard to automate; it depends on interpreting the perspective (concept) ‘family background’ as an instruction to pick out a skeleton, following kinship relations, and then elaborate each entity in the skeleton, following personal-history relations.

Above we define three methods to select knowledge from a knowledge base. All three methods start from a focused entity and navigate inside the knowledge base to define the required information. The navigation plan is what we called perspective here. Therefore, an extended topic is a focused entity plus a perspective, as shown in figure 3-4. We may notice that both of method 2 and method 3 include a general concept that confines the required information, which will be referred to as the generic part of the topic expression, and the rest of the question refers to specific relationship or entity in the knowledge base, we call them the specific part. We notice that since the generic part confine the required information, therefore it will be replaced in the answer by details that are unknown to the questioner, while the specific part will be kept as the *given* in the answer. Thus, we say that an extended topic may consist of two parts, a generic part pointing to the unknown and a specific part pointing to the known. The first part is equivalent to the given in the discourse and the second part is a sketch of the new information in the discourse, as shown in the following chart. To simplify the name, we will use terms ‘generic topic’ and ‘specific topic’ to refer to the generic part and the specific part of extended topic respectively.

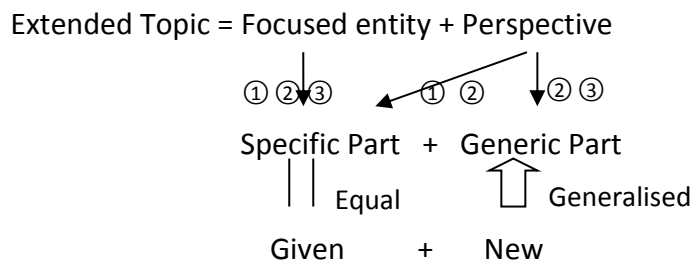


FIGURE 3-4. COMPOSITION OF EXTENDED TOPIC

Back to the first example, when we ask for what mammal eats grass. One answer could be ‘cows and sheep eat grass’. We see here that mammal in the question is replaced by a specific kind of mammal – i.e., cows and sheep; while ‘eat grass’, as part of the known knowledge, is

kept in its original form in the answer. Here connects back to the observation given in section 2.3.3, where we showed that an indicative abstract contains phrases that refer to the comments in a corresponding informative abstract. For example, the phrase ‘advertising budget’ mentioned in the indicative abstract relates to ‘will spend 15 million sterling on advertising’ in the informative abstract. In chapter 1 and chapter 2, we cited Boguraev and Kennedy’s (1997) work on topic generation. Below we will again use an example passage ([3.2]) extracted from this work. According to the author, the passage below talks about “the background of Amelio”. We might generalise that ‘53’ is about his age and ‘rescue of National Semiconductor from near-bankruptcy and 16 patents, including one for co-inventing the charge-coupled device’ is about his deeds, both of which are a particular aspect of background.

[3.2] Amelio, 53, brings a lot of credibility to this task. His resume includes both a rescue of National Semiconductor from near-bankruptcy and 16 patents, including one for co-inventing the charge-coupled device.

Above we see that extended topic could be derived from the discourse by selecting prominent entities and by the operation of generalisation. These operations are quite similar to the macrorules defined in van Dijk (1977) (refer to the section 2.3.1.2 for more details). However, the topic expression generated using van Dijk’s methods are informative topics. The difference lies in that: a) from functional point of view, the concept generalised in van Dijk’s (1977) defined process should “dominate” all the details in the discourse; in this sense, it does not lose any information; while the generic part of extended topic would just use some general concepts known to most ordinary people (or at least known to a typical audience of the discourse) instead of conveying new relationships; b) typical general concepts used in extended topic are category names (e.g., animal) and attribute names (e.g., colour), while typical concepts in van Dijk’s notion of topic are predicates (e.g., rising up) or values of attributes. For the example topic generated by van Dijk given in section 2.3.1.2, “A (little) town (called Fairview) is declining because it cannot compete with another town (called Bentonville)”, the corresponding extended topic would be “the economic status of Fairview”.

There are two further points we would like to make. First, the focused entity used above is not necessarily a named entity. For example, a piece of fact such as ‘John quarrels with his father’ could also be seen as a focused entity; and an example extended topic consisting of this focused entity is ‘the reason why that John quarrels with his father’. Second, it is worth noting that the specific topic and the generic topic are not named by a direct comparison with each

other. As noted before, the generic topic defines a general characteristic of the required information and will be replaced in the discourse by elaborated details; while specific topic points to something that are already known and will be kept in the discourse. In other words, generic topic and specific topic are two different roles. In fact, we could find a general concept taking the role of the specific topic. For example, in topic ‘the origin of *animals*’, ‘animals’ is the specific topic. Further, given that a discourse is talking about a general concept such as ‘biography’, people would have different predictions of the discourse content depending on which role they presume the concept plays. When ‘biography’ is the generic topic, people would expect to see that the discourse contains biographical facts of a particular person (e.g., his education and career development); when it is the specific topic, people would expect to see that the discourse talks about some general attributes of biography as a type of written work (e.g., key facts in a typical biography). In appendix E we provide many example discourses to illustrate the difference between the two roles. Further study on the two roles is out of the scope of this thesis.

3.3 CONTEXTS IN WHICH EXTENDED TOPIC ARE EXPLICITLY FORMULATED

We have noted that the generic part of an extended topic is generalised from the discourse. Therefore, the extended topic of a discourse might not be explicitly expressed in the discourse but could still be inferred by the reader. This does not mean that we cannot find places where extended topic is explicitly formulated. In previous sections, we have given examples of indicative summaries (example [2.5]) and discourse anaphora (example [1.4]) which contain extended topics. In this section, we will discuss a set of such contexts, including:

- WH-questions,
- titles and section headings,
- sentences describing the plan of a document,
- anaphoric references to a discourse unit,
- entries in an index,
- indicative summaries/abstracts.

WH-QUESTIONS

Probably WH-questions are the most reliable source of extended topic. A WH-question can be thought of as an information retrieval device, through which the questioner indicates but not tells the desired information. Most WH-questions can be recast (if necessary) into the form ‘What is/are the *G* of/for/that *S*’. Once a question is recast into this form, we get a nominal phrase¹³ – the typical syntactic realisation of an extended topic.

[3.3] How much does a packet of aspirin cost?

What is the price of a packet of aspirin?

[3.4] Why should a course of anti-Biotics be completed?

What is the benefit of completing a course of anti-Biotics?

Most of the time, *G* corresponds to the generic part and *S* corresponds to the specific part in the structure of extended topic. In the linguistic study of the structure of WH-questions, it is established that the question word denotes the focus and will be replaced by the new information in the answer. In the form given above, since *G* and the question word are linked by a copula, the concept in *G* should also refer to the new information in the answer. This explains why *G* should be the generic part. We describe in the next chapter an analysis of some medical FAQs showing that most questions fit into this pattern.

TITLES AND SECTION HEADINGS

Many titles or section headings fit into the patterns we have suggested for extended topic, such as [3.5] and [3.6].

[3.5] Method and Theory in Historical Archaeology¹⁴

[3.6] The Life of Samuel Johnson¹⁵

However, our impression is that functions of titles vary across different types of discourses. For example, titles of newspaper articles play the role of describing the key event and enticing readers to read the details. [3.7] and [3.8] are two headlines obtained from Yahoo¹⁶.

[3.7] AP Poll finds Bush, Kerry locked in a tie

¹³ In section 2.3.3, we note that an indicative topic expression should take the form of a nominal phrase

¹⁴ South, S. (1978) *Method and Theory in Historical Archaeology*. Academic Press, New York, New York.

¹⁵ Boswell, J. (1952) *The Life of Samuel Johnson*. Modern Library: New York.

¹⁶ Refer to <http://news.yahoo.com/>

[3.8] Isreal kills top Hamas militant ahead of Gaza vote

In general, we believe that titles tend to be informative. Example [3.9] is the title of an academic paper. Example [3.10] is a topic expression used to describe the plan of the same document. They are taken from Spangler et al. (2002)¹⁷. We see that example [3.10] only tells us that the paper is about a system and a methodology; while in example [3.9] the author directly informs the name of the system and the concrete method being used.

[3.9] mindmap: utilizing multiple taxonomy and visualization to understand a document collection.

[3.10] We present a novel system and methodology for browsing and exploring topics and concepts within a document collection.

SENTENCES DESCRIBING THE PLAN OF A DOCUMENT

As example [3.10] illustrates, topic expressions used to describe the plan of a document have the nature of an extended topic. The opening paragraph of a scientific paper often contains a sentence of the form ‘This paper describes/presents a G of S’ where G is the generic part of an extended topic and S is the specific part. Towards the end of the introduction there may be a paragraph on the plan of the paper, giving the topic of each section. Below are examples taken from two scientific papers found through a web search.

[3.11] Section 3 presents the design of a remote memory pager.

[3.12] This paper presents evidence of self-similarity in WWW traffic.

We describe in chapter 4 a study showing that this is a common pattern, and a useful source of concepts like ‘theory’ and ‘design’ which denote a perspective applicable to a wide range of domains.

ANAPHORIC REFERENCES TO DISCOURSE UNIT

When we refer back to the content of a whole paragraph or section, we often do so in terms of its perspective rather than its focused entity. References forward, as we have seen, often have the form ‘the P of E’; references back are more likely to mention only P (e.g., ‘this explanation’, ‘our theory’, ‘the algorithm’). Such references are examples of what is sometimes called ‘discourse deixis’, as distinct from phrases like ‘this section’, ‘the previous

¹⁷ An academic paper in computational linguistics

paragraph’, are examples of “document deixis” (Paraboni, 2003) because they apply to the text itself rather than its content.

ENTRIES IN AN INDEX

The analytical index at the end of a book often has two levels when an entry is complex. The first level is more likely to refer to a focused entity, such as a person or place; the second level often mentions perspectives from which the person or place is approached. Here is an example from a book on British History (*The Isles* by Norman Davies)¹⁸:

[3.13] Elizabeth I, Queen of England and Ireland: accession, 392; birth, 388, 432; death, 451; heir, 378, 411, 452, 455; historical verdict on, 437 imprisonment and execution of Mary Queen of Scots, 380, 411; knights Drake, 404; Parliament, 384; personality, 426, 438; relationship with Spain, 398; religion, 392, 437.

In the above example, the first level presents the focused entity—Elizabeth—and the second level lists various perspectives from which the focused entity is addressed.

INDICATIVE SUMMARIES/ABSTRACTS

In example [2.5], we show that many topic expressions in the indicative abstract contain both the topic element and a general characterisation of the comment, which is the typical structure of extended topic. Further, the structure of extended topic is also evident in some automatic summary generation system. Kan et al. (2001) presents CENTRIFUSER which generates summaries for highly structured documents. The generated summary contains both informative elements and indicative elements. Specifically, he uses the indicative part to summarise the content difference between different documents. He uses a topic tree to represent the structure of documents so as to compare the differences. A sample topic tree he provides contain a node “angina” at a high level and perspectives such as “causes”, “symptoms”, “diagnosis”, “treatment” underneath.

3.4 OUTLINE OF THE THEORY

We now summarise the ground that has been covered. We will provide empirical experiments on topic expressions to validate the theory in chapter 4 and list the potential applications of the theory in chapter 5.

¹⁸ *The Isles, A History* by Norman Davies, Oxford University Press, 1999

WHAT IS EXTENDED TOPIC?

An extended topic is a concise, indicative formulation of the content of a discourse. By ‘concise’ we mean that it should be expressible in a short phrase. By ‘indicative’ we mean that it should not reproduce the informational content of the discourse: in other words, it should represent the question that the discourse addresses without giving the answer.

WHY IS EXTENDED TOPIC IMPORTANT?

By concisely expressing a notion of relevant content, it provides a convenient input for programs that perform information retrieval or document generation. It also helps to analyse or generate titles, WH-questions, discourse-deictic references, and entries for an analytical index.

HOW CAN THE NOTION BE MADE PRECISE?

Since extended topic provides a criterion for relevant content, a formal definition will depend on details of the knowledge representation. We have assumed here a simple general model in which knowledge comprises a network of entities and relationships. In terms of this model, we can define an extended topic as a function which, given a knowledge network, identifies a relevant subnetwork (or perhaps distinguishes degrees of relevance).

WHAT ARE THE COMPONENTS OF AN EXTENDED TOPIC?

We suggest that an extended topic identifies a relevant subnetwork by the following strategy: first, pick an entity (or entities); second, apply a rule for navigating from this entity. The components of an extended topic are accordingly the focused entity (which identifies the entity from which we navigate) and the perspective (which identifies the navigation rule). It is this second component, the perspective, which is insufficiently addressed in previous work on topic.

HOW CAN PERSPECTIVE BE DEFINED?

The perspective allows retrieval of entities and facts by defining their relationship to the focused entity. The simplest kind of perspective is therefore a single relation, like price, which would retrieve an amount of money if the focused entity was a product or service (e.g., ‘What is the cost of a DNA test?’). A slightly more complicated kind of perspective contains a single relation and a general category that defines the type of thing having the particular relation with the focused entity. Both of the two kinds of perspectives would retrieve a single entity or a single piece of fact. To retrieve material for longer discourses, a more complex relational

Chapter 3 – Theory of Extended Topic

definition is needed. Words like ‘procedure’, ‘explanation’ and ‘proof’, can be viewed as labels for rather complicated navigation plans for retrieving subnetworks around a focused entity. In part, they work by privileging a particular set of relevant relations.

HOW DO THE COMPONENTS OF AN EXTENDED TOPIC RELATE TO COMPONENTS IN A RELEVANT DISCOURSE?

The focused entity together with the single relationship in the first two kinds of perspectives constitute the specific part of an extended topic, the general category in the second kind of perspective and the complex navigation plan in the third kind of perspective are defined as the generic part. In a relevant discourse, the specific part will be kept as the given information and the generic part will be replaced by the new information.

CHAPTER 4

PROBING THE STRUCTURE OF EXTENDED TOPIC

4.1 INTRODUCTION

In the last chapter, we have suggested that a topic formulation is rather like a query to a knowledge base: given a specific knowledge base, it determines which parts are (more or less) relevant. We have also shown how the two elements of an extended topic, focused entity and perspective, combine to perform this task. First, the focused entity selects a node in the knowledge base. Next, the perspective tells us how to navigate out from this node to select a relevant region. There are different kinds of perspectives. The simplest kind of perspective pinpoints a single relation in the knowledge base to retrieve another entity that holds the specified relation to the focused entity. More complicated kinds of perspectives would contain a general concept which specifies as relevant a family of entities or relationships. We defined the general concept as the generic part of extended topic and the rest (the focused entity and the single relation) as the specific part. Further, we suggested that the generic part could be derived from the new info from a relevant discourse and the specific part corresponds to the old information. In this chapter, we will introduce several empirical studies to support the psychological and linguistic validity of extended topic.

In section 4.2, we study WH-questions and show that most of them could be recast into the form of 'what is/are the G of/that/for S'. We acquire a set of generic topics by extracting the G parts and further classify them into different types. In section 4.3, we design an experiment with human subjects to test whether people could recognise the generic topic of a discourse. In section 4.4, we collect a set of phrases describing the plan of academic papers and separate them into the generic part and the specific part based on simple linguistic patterns. We compare between these two parts to verify whether the concepts used in the generic parts are more general in comparison to those in the specific parts. In section 4.5, we study the correlation between the generic/specific part of an extended topic and the given/new element in a related discourse.

4.2 WH-QUESTION ANALYSIS

To get a sense of the range of generic concepts¹⁹ we perform a small scale study on WH-questions. The generic part is often implicit in a WH-question, but we can get the generic part by simply recasting the question into the form of ‘What is/are the *G* of/for/that *S*’. Below are some examples of how the questions are analysed.

[4.1] Example WH-Questions:

Original question How much does a DNA test cost?

Revised question What is the cost of a DNA test?

Generic Part Cost

Original question What are the side-effects of protein supplements?

Generic Part Side-effects

Original question What is Cryptosporidium?

Revised question What is the definition of ‘Cryptosporidium’?

Generic Part Definition

Original question Why do people commit suicide?

Revised question What is the cause of suicide?

Generic Part Cause

We have done the analysis on different sets of WH-questions. One set contains 77 frequently asked questions (FAQs) in the medical domain. We asked two annotators to recast the questions into the above-mentioned form and then compared their results to see whether they derive the same generic part. We found that only 69 questions were properly transformed by the second annotator, among which 34 have the same generic part as suggested by the first annotator. We further examined why they got different results for the rest of the questions. One reason is that they choose different terms which have similar meanings. For example, for question ‘what causes human to become thirsty?’, one annotator transforms it to be ‘what is he reason that human become thirsty’ but the other one uses ‘cause’ instead of ‘reason’. Similarly, for question ‘how can you cope with fibromyalgia?’, one annotator transforms it to be ‘what is the best way to cope with fibromyalgia’ but the other

¹⁹ General concepts typically used in the generic part of extended topic

Chapter 4 – Probing the Structure of Extended Topic

one uses ‘method’ instead of ‘way’. Another reason is that one annotator chooses a more specific term than another. For example, for question ‘who takes medication?’, one annotator transforms it to ‘what are the types of people that take medication?’ but the other one uses term ‘recipient’ instead of ‘types of people’. The questions and the annotation results are included in appendix A.1.

We repeated the same experiments on questions used in TREC²⁰. One question set contains 60 questions used in the question answering track in TREC 2004²¹. The results of the two annotators match for 36 questions and do not match for 24 questions. Refer to appendix A.2 for the details. Another question set contains 5 questions used in the ciQA task in TREC 2007²². The two annotators have the same result for all of the 5 questions. Details are included in appendix A.3.

We further classify the concepts in the generic part of the medical domain question set as follows.

Group	Generic concept	Frequency
Causal	Cause, Effect	16
Procedural	Condition, Method, Response, procedure, Purpose, Motive, Constraint, Management, Instruction	15
Inferential	Reason, Risk, Result, Diagnosis, Explanation, Interpretation, Symptom	13
Conceptual	Definition, Meaning, Importance, Type, Relationship	12
Temporal	Time until, Duration	6
Roles	Recipient, User, Experience, Happening	4
Instrumental	Cure, Treatment	4
Properties	Amount, Cost, Colour, Size	4
Spatial	Locus, Composition	3

TABLE 4-1. A CLASSIFICATION OF GENERIC CONCEPTS

²⁰ <http://trec.nist.gov/>

²¹ http://trec.nist.gov/data/qa/t2004_qadata.html

²² <http://www.umiacs.umd.edu/~jimmylin/ciqa/>

Chapter 4 – Probing the Structure of Extended Topic

The above generic concepts all have a relational character, but they differ in complexity, and hence in the amount of information requested. This difference can be made more precise by comparing the questions with queries to a relational database. On a rough classification we could distinguish four cases:

1. Retrieving the value of a simple attribute. The Property questions (amount, cost, colour, size) belong in this category.
2. Retrieving an entity that participates in a relationship. Questions about Role (e.g., 'Who discovered penicillin') are examples.
3. Retrieving an event or a proposition by its relationship with other events or propositions. For instance, a purpose question like 'What is the purpose of aspirin' might elicit a single proposition such as 'Aspirin is used for treating headache'. Or a Definition question like 'What is the pelvic wall' might elicit a single fact.
4. Retrieving a network of related events or propositions. The relationships in the network might be linked by planning relations (e.g., 'How should I prepare the apparatus'), or by cause and effect relations ('What are the consequences of a stroke?'), or by inferential relations ('What is the evidence that aspirin is safe?'), or by spatial relations ('What is the structure of the pelvis?'), and so forth.

Roughly, the typical answers to these types of query would be (1) adjective, (2) noun phrase, (3) single sentence, and (4) paragraph. For the first three types, the navigation process would normally be straightforward: it suffices to follow a single arc from the focused entity. The hard case is the fourth type. How can a generic concept serve as a criterion for retrieving a complex sub-network? Yet it is precisely these complex concepts that are most important in discourse planning and information retrieval. In general terms, it seems clear that the answer will lie in some kind of navigation from the focused entity (or entities) based on relations. We believe that complex generic concepts like 'procedure', 'explanation', 'career', serve as a shorthand for families of relations: for instance, 'procedure' implies relations like goal, precondition, instrument, method; 'explanation' implies cause, motive, and so forth. However, as we have seen in the example of Mozart's family background, the navigation process is not merely one of following relevant relations as far as possible; we need to consider how much material to embrace, and whether there are secondary relations that should be followed once the skeleton of a discourse has been identified.

4.3 EXPERIMENT I

4.3.1 INTRODUCTION

In section 4.2, we claimed that some generic concepts work as a shorthand for a family of relations. Therefore, given a generic concept, we are able to judge whether a passage is relevant according to whether it contains facts belonging to the family. Note that a specific concept such as ‘Mozart’ also sets a criterion for selecting relevant facts. However, this criterion can be simplified as whether ‘Mozart’ is part of the fact. The relevancy defining function of a generic concept is different. It is the aim of this experiment to investigate this function. Specifically, we will show that subjects are able to identify relevant passages even they do not contain the concept explicitly. In section 4.3.2, we will present the experiment methodology; in section 4.3.3, we will describe in details the data preparation process; the experiment result is presented in section 4.3.4; we draw conclusions in section 4.3.5.

4.3.2 METHODOLOGY

Directly asking subjects whether a given passage is relevant to a given generic concept may confuse them since there is no clear border between relevant facts and irrelevant ones. We instead give subjects two passages at a time and ask which one is more relevant to the given concept. We then examine the patterns in subjects’ choices. Our hypotheses are: a) when one passage is clearly more relevant than the other one, subjects are able to pick up the most relevant one; b) when both passages are relevant or neither passages is relevant, subjects tends to select one randomly.

4.3.3 DATA PREPARATION

We selected 7 generic concepts from the set collected in the WH-question analysis experiment. The criterion was to choose, from each category, one most frequent concept that defines a network of related events or propositions (e.g., ‘composition’) rather than a single value, entity or proposition (e.g., ‘colour’). The selected concepts are given in the table below.

For each generic concept, we collected three relevant and three irrelevant passages from the Web. To collect relevant passages for a concept, we searched the concept in Google and extracted passages with the concept explicitly included or implied in its title (implied based on our own judgement). Neither the relevant passages nor the irrelevant ones contain the concept explicitly. Refer to appendix B.1 for all the passages. Note that we do not take this

relevant/irrelevant annotation as the ground truth. We just use it as a reference for generating the questionnaires.

Concept
Cause
Response
Reason
Definition
Experience
Treatment
Composition

TABLE 4-2. THE SELECTED GENERIC CONCEPTS

We then generated 30 ordered passage pairs for each concept by exhausting all the possible permutations of the six passages. We randomly distributed these 30 passage pairs into 30 questionnaires. Therefore, each questionnaire would contain seven ordered passage pairs, one for each concept. Our questionnaire generation method is to ensure a good level of randomness. One sample questionnaire is included in appendix B.2.

4.3.4 RESULT

We asked 60 subjects to do the experiment, 2 for each questionnaire. Refer to appendix B.2 for the detailed instructions. The time subjects spent ranged from 5 minutes to 15 minutes. The level of agreement between subjects in their choices is described as follows.

Among the 210 questions, 126 are to choose between a relevant passage and an irrelevant one (the first set), 84 are to choose between either two relevant passages or two irrelevant ones (the second set). We have two answers for each question. For the first set, there are 94 agreements (out of 126); in comparison, for the second set, there are only 49 agreements (out of 84). The Chi Square statistics shows that this difference is significant (Chi-Square=6.141; $p \leq 0.025$). Refer to table 4-3 for the contingency table. We also compare the above result in the second set with random distribution and found there is no significant difference (Chi-Square=0.386; $p > 0.05$). Refer to table 4-4 for the contingency table.

Besides, among the 252 answers for the questions in the first set, 214 agree with our expectations. The difference between this distribution and a random one (half agree and half disagree) is significant (Chi-Square= 69.996; $p \leq 0.0001$). Refer to table 4 for the contingency table.

Chapter 4 – Probing the Structure of Extended Topic

Moreover, among the whole answer set, 219 choose the first passage, 201 choose the second one. Therefore, the order of passages has no significant influence on the decision.

	Num of agreements	Num of disagreements
Set 1	94	32
Set 2	49	35

TABLE 4-3. COMPARE THE LEVEL OF AGREEMENTS BETWEEN SET 1 AND SET 2

	Num of agreements	Num of disagreements
Set 2	49	35
Random	47	47

TABLE 4-4. COMPARE THE LEVEL OF AGREEMENTS BETWEEN SET 2 AND A RANDOM DISTRIBUTION

	Num of agreements	Num of disagreements
Experiment Result	214	38
Random	126	126

TABLE 4-5. EVALUATE THE LEVEL OF AGREEMENTS BETWEEN SUBJECTS' CHOICES IN SET 1 AND THE EXPECTED RESULT

4.3.5 CONCLUSION

Presuming that our initial relevance annotations are correct, the experiment results confirm with our hypotheses presented in section 4.3.2. Even without this presumption, the strong agreement between human subjects in the first set of result suggests human subjects do have a shared view on which passage is more relevant to a generic concept. Furthermore, the agreement between subjects' choice and our initial annotation suggests that our perception matches to what the subjects perceive. Besides, the high level of disagreement in the second set of results removes interfering factors such as that some subjects may copy others' answers.

4.4 EXPERIMENT II

4.4.1 INTRODUCTION

As stated in chapter 3, extended topic is expressed explicitly in phrases describing the plan or the purpose of scientific papers. It is the aim of this experiment to further examine the nature of such topic expressions. Specifically, we will compile a collection of topic expressions and design experiments to test whether the structure of such topic expressions shows the same pattern as extended topic. Based on the collected topic expressions, we will also compile a list of typical concepts used in the generic part of extended topics.

As stated in section 3.3, such topic expressions often take the form of ‘This paper describes/presents the/a G of/that/for S’, where G is the generic part and S is the specific part, as it is in example [4.2].

[4.2] This paper describes a novel computer-aided procedure for generating multiple-choice tests from electronic instructional documents.

However, this pattern does not always hold. For example, in sentence [4.3] the phrase ‘ontology-based’ belongs to the first noun phrase and is therefore considered as the generic part; in sentence [4.4], which is semantically equivalent to sentence [4.3], the same phrase is regarded as the specific part. Since in principle all modifiers in a noun phrase can be post-posed (to the specific part position), only the head noun in a noun phrase is bound to be the generic part. We therefore propose that all elements other than the head noun are regarded as the specific part.

[4.3] This section presents an ontology-based framework for linguistic annotation.

[4.4] This paper presents a framework for linguistic annotation that is ontology-based.

As mentioned in last chapter, the generic part and the specific part of an extended topic are not named by a direct comparison between them. Therefore, the generic part of an extended topic is not necessarily more general than the specific part of an extended topic. Nonetheless, we believe that the terms used in the generic part are usually more general than the terms in the specific part. If this is proved to be true with our collected topic expressions, it will at least indicate that such topic expressions do have a significant internal structure.

Previous studies (Justeson and Katz, 1995; Boguraev and Kennedy, 1997) have shown that nouns are the main content bearer, and our experiment only examines the nouns in the topic expressions. A preliminary hypothesis is formulated as follows.

General Hypothesis: Among the nouns in the phrases describing the plan or the purpose of scientific papers (excluding those in the leading cue phrase such as ‘this paper describes’), head nouns are more general than non-head nouns.

The next question is how to measure the generality of a noun. One clue is that general term types should be less than specific term types. This means, if we collect the same number of general term tokens and of specific term tokens and put them into two groups, the first group

should contain fewer unique terms (i.e., term types). In general, the distribution of general terms should be more focused (or narrower) than the distribution of specific terms. Figure 4-1 shows the difference between narrow distribution and wide distributions. The X-axis of the figure represents the rank of terms based on frequencies, and the Y-axis represents the term frequency. We can see that a certain proportion of top ranked terms tend to take a larger percentage of term frequencies in a narrow distribution than in a wide distribution.

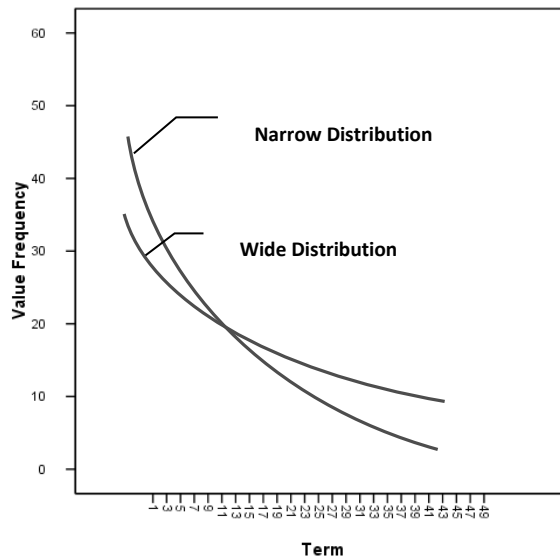


FIGURE 4-1. NARROW VS. WIDE TERM DISTRIBUTIONS

The original hypothesis can be reformulated as below.

Hypothesis I: If all the nouns in a collection of phrases describing the plan or the purpose of scientific papers (excluding those in the leading cue phrase such as ‘this paper describes’) are extracted and separated into head nouns and non-head nouns, the distribution of head nouns is narrower than that of non-head nouns.

We also measure using human judges. However, it is confusing to ask whether a term is a general or how general a term is. First, there is no a clear boundary between general concept and specific concept; instead, there might be degrees of generality. Moreover, two randomly selected concepts might not be comparable to each other. For example, we may say that concept ‘colour’ is more general than concept ‘red’, but we are not able to compare between ‘colour’ and the concept ‘cow’. In our experiment, we collected topic expressions from academic papers in both physics and computational linguistics. Instead of asking whether a term is general or specific, the choices we provided are ‘term specific to a particular research subject’ or ‘term applied to scientific research in general’. The metric of scientific-general and scientific-specific might not overlap with the metric of general or specific in our sense.

However, we believe that these two metrics are strongly correlated. One point is, many scientific specific terms such as ‘DNA101’ in biology refers to a concrete entity. The hypothesis can be adapted as follows.

Hypothesis II: If all the nouns in a collection of phrases describing the plan or the purpose of scientific papers (except those in the leading cue phrase such as ‘this paper describes’) are extracted and separated into head nouns and non-head nouns, human judges will rate the head noun group as containing more scientific-general terms than the non-head noun group.

The above analysis focuses on noun phrases describing the plan or the purpose of scientific papers. One extended study is to verify whether it is generally true that head nouns are more general than non-head nouns in noun phrases. To answer this question, we will also experiment on a set of noun phrases and compare the head nouns and the non-head nouns that they contain. The details will be presented in section 4.4.5.

4.4.2 METHOD

4.4.2.1 DATA ACQUISITION

TOPIC EXPRESSIONS ACQUISITION

Topic expressions were taken from articles in the Institute of Physics²³ and ACL Anthology²⁴. Both provide a searching facility on their websites. From these two websites, we collected topic expressions following the cue phrases ‘this paper presents’, ‘we present’ and ‘this paper describes’, etc. From the website of the Institute of Physics, by searching ‘this paper presents’, the system retrieved 309 text snippets, from which 229 topic expressions were extracted. On the same website, by searching ‘this paper describes’, we acquired 207 topic expressions extracted from the 257 initially retrieved text snippets. We repeated the same process on the ACL Anthology website. By searching key words ‘section’ and ‘presents’ together, we acquired 378 relevant documents. From these documents, we extracted 435 valid sentences by matching pattern ‘section+<NUM>+presents’. On the same website, we also extracted 456 topic expressions by matching ‘we describe’, ‘this paper describes’ and ‘section+<NUM>+describes’. The discussion about the experiment results will always be divided into four categories: ‘CL+Present’, ‘CL+Describe’, ‘Physics+Present’ and

²³ Refer to <http://www.iop.org>

²⁴ Refer to <http://acl.ldc.upenn.edu>; ACL Anthology provides a collection of academic papers in computational linguistics

‘Physics+Describe’. In appendix C.1, we provide some sample sentences for each of these categories. The table below summarises the number of topic expressions collected for each category.

Category	Number of Topic Expressions
CL+Describe	456
CL+Present	435
Physics+Describe	207
Physics+Present	229

TABLE 4-6. NUMBER OF COLLECTED TOPIC EXPRESSIONS

BUILDING HEAD NOUN AND NON-HEAD NOUN LISTS

We expected that some simple heuristics based on POS tags would allow us to extract head nouns and non-head nouns from topic expressions. The POS tagging tool we used is called NLProcessor²⁵. Based on POS tags, we used the following algorithm for finding the head noun of this expression:

In most cases, we find the last noun in the first noun group. This method works for the following examples.

[4.5] This paper presents detailed calculations of the flow.

[4.6] This paper presents a novel fabrication process for a tapered hollow metallic microneedle array.

In some cases, there are two or three parallel head nouns. In this case, we check the constituent that directly follows the first head noun; if it is a parallel structure marker, such as ‘,’ and ‘and’, we will continue to find the last noun of the first noun group in the phrase led by these markers. This algorithm covers the following example:

[4.7] This paper presents the design and experimental results of control of an SMA actuator using pulse width modulation.

Due to the ambiguity of parallel structure, we only allow definite articles and adjectives to be in between of the parallel structure marker and the second head noun. This requirement removes concept ‘expression’ in the following sentence that might be wrongly categorised otherwise.

²⁵ Refer to <http://www.infogistics.com/posdemo.HTML>

[4.8] This paper presents an analytical method and obtains an analytical expression.

The java implementation of the algorithm can be found in appendix C.2. From the above description, we can see that this algorithm cannot deal with some complex sentence structures. For example, it cannot deal with distant parallel components, such as ‘process’ and ‘solutions’ in the following example:

[4.9] This paper presents a novel fabrication process for a tapered hollow metallic micro-needle array using backside exposure of SU-8, and analytic solutions of critical buckling of a tapered hollow microneedle.

We have done an empirical study of the effectiveness of the algorithm in locating head nouns. We randomly chose 35 topic expressions which contain 41 head nouns. Using this algorithm 38 head nouns are correctly located, two head nouns are missing, one is extracted wrongly. The precision is 97.4%, and the recall is 92.7%. The sentences and the head nouns the system found are included in appendix C.3.

We also found that there were some topic expressions that do not take the presupposed form of an extended topic. For example, in sentence [4.10], the object of the main predicate is not a noun phrase; in sentence [4.11] and [4.12], noun phrases “CREAM” and “the classifier algorithm” do not have any following modifier, which are named by us as bare NPs. The first kind of topic expressions rarely appears, while bare NPs take a small proportion. Of all the topic expressions in category ‘CL+describes’, we found 34 bare NPs; of all the topic expressions in category ‘Physics+presents’, we found 4 bare NPs. In our experiment, we ignored the existence of these two kinds of topic expressions.

[4.10] This section 6 describes how the method was generalised to cover other semantic features.

[4.11] Section 4 presents CREAM.

[4.12] This Section 4 describes the classifier algorithm.

The extracted head nouns and non-head nouns were put into two lists, each of which contains a number of terms ordered by their frequencies. We also designed a simple algorithm for lemmatising nouns so that the plural form and the singular form of a term can be combined.

4.4.2.2 TERM GENERALITY JUDGEMENT BY HUMAN SUBJECTS

MATERIAL

The above process generated two term lists for each category. There are over 1000 terms in all the lists of the four categories. It is not realistic to give all the terms to a human subject for him/her to judge whether they are scientific-research-general or research-domain-specific. We used a subset of terms to represent the whole population. Specifically, we chose the most frequent terms to ensure a small number of terms and a better coverage. The list of representatives was compiled by selecting terms from each category. Since the topic expressions obtained for computational linguistics are more than those for Physics, the accumulated term list (104 terms) contains more CL terms (around 75 terms) than Physics terms (around 45 terms).

SUBJECTS

Three students attend the experiment, two specialising in computational linguistics (subject 1 and subject 2), and one with both computational linguistics and physics background (subject 3).

PROCEDURES

For each term within the list, the subjects are required to make a choice from the following three categories: 'computational linguistics term', 'physics term' and 'general scientific term'. Refer to appendix C.5 for the questionnaire.

4.4.2.3 HEAD NOUN AND NON-HEAD NOUN COMPARISON

DISTRIBUTION TEST

The terms in our acquired head noun lists and non-head-noun lists were ordered by term frequencies, as illustrated in figure 4-1. In section 4.4.1, we supposed that narrowness of term distribution could be reflected by the percentage of term frequencies taken by a certain proportion of terms at the front of the list. We compared the two populations by the percentage of frequencies taken by the 5% top ranked terms and %10 top ranked terms. This comparison is for testing hypothesis I.

PROPORTION OF GENERAL TERMS TEST

After the term-generality judgement experiment, we obtained a list of terms judged as either scientific-research-general or research-domain-specific. This list contains terms from four

categories. This means that in each category we have got some judged terms that can be used for further test. We then compared the proportion of general terms between the head noun group and the non-head noun group within each category. This comparison is for testing hypothesis II.

4.4.3 RESULT

4.4.3.1 RESULT OF TERM DISTRIBUTION EXPERIMENT

The results of the term distribution experiment are as follows. Under each category, the first table shows the number of terms and the term frequencies in the head noun list and the non-head noun list; the second table shows the percentage of term frequencies taken by a certain proportion of terms (5% and 10%). We include in appendix C.4 the top ten percents head nouns and top five percents non-head nouns of each category.

Category	Group	Number of Terms	Term Frequency
CL+Describe	Head Noun	163	470
	Non-Head Noun	573	1622
CL+Present	Head Noun	128	473
	Non-Head Noun	374	1262
Physics+Describe	Head Noun	66	148
	Non-Head Noun	405	673
Physics+Present	Head Noun	90	256
	Non-Head Noun	607	1244

TABLE 4-7. NUMBER OF TERMS AND TERM FREQUENCIES

Category	Group	5% Term	10% Term
CL+Describe	Head Noun	37.23%	49.69%
	Non-Head Noun	30.85%	46.91%
CL+Present	Head Noun	41.31%	52.34%
	Non-Head Noun	31.01%	37.32%
Physics+Describe	Head Noun	31.82%	42.57%
	Non-Head Noun	22.29%	40.50%
Physics+Present	Head Noun	32.03%	48.83%
	Non-Head Noun	25.24%	36.74%

TABLE 4-8. TERM DISTRIBUTIONS

4.4.3.2 RESULT OF THE GENERAL VS. SPECIFIC SCIENTIFIC TERM EXPERIMENT

We used Cohen’s (1960) version of the kappa statistic to measure agreement among human subjects. The table below gives the kappa values. The average value of agreement between pairs of subjects is 0.4243. This kappa value indicates moderate agreement.

	Subject1	Subject2	Subject3
Subject1	-	.404	.502
Subject2	.404	-	.367
Subject3	.502	.367	-

TABLE 4-9. AGREEMENT BETWEEN TWO SUBJECTS BASED ON KAPPA STATISTIC

Refer to appendix C.6 for the experiment result and the formulae for calculating the kappa value.

There are 100 terms on which at least two subjects agree in their judgement. We used these terms for testing the proportion of general terms. This list of terms and the associated judgements are included in appendix C.7.

4.4.3.3 RESULT OF TESTING THE PROPORTION OF GENERAL SCIENTIFIC TERM

We compared the proportion of general terms between the head noun list and the non-head noun list (limited to 100 representative terms). Some basic statistics of the representatives and the comparison results are shown below. Table 4-10 shows the number of representative terms, their frequencies and the proportion of them over the whole population; table 4-11 contains the frequencies of scientific-research-general terms and research-domain-specific terms in both head noun list and non-head noun list; table 4-12 shows the chi-square values of the difference. The chi-square values of all the four categories suggest that the head noun group contains a significant larger proportion of scientific-research-general terms than the non-head noun group.

Brief inspection of these statistics indicated no important difference between the topic expressions following ‘present’ and those following ‘describe’. We therefore combined the data of category ‘present’ and the data of category ‘describe’ within each domain together, as shown in table 4-13. The chi-square values are 235.877 [p<0.0001] for CL and 82.317 [p<0.0001] for physics.

Chapter 4 – Probing the Structure of Extended Topic

Category	Group	Number of Terms	Term Frequency	Proportion of Terms	Proportion of Term Frequencies
CL+Describe	Head Noun	28	267	16.8%	56.8%
	Non-Head Noun	35	580	6.1%	35.8%
CL+Present	Head Noun	30	319	23.4%	67.4%
	Non-Head Noun	43	507	11.5%	40.2%
Physics+Describe	Head Noun	22	96	33.3%	64.9%
	Non-Head Noun	26	159	6.4%	23.6%
Physics+Present	Head Noun	34	188	37.8%	73.4%
	Non-Head Noun	34	285	5.6%	22.9%

TABLE 4-10. STATISTICS OF THE REPRESENTATIVE TERMS

Category	Group	Frequency of Scientific-Research-General Terms	Frequency of Research-Domain-Specific Terms
CL+Describe	Head Noun	228	39
	Non-Head Noun	302	278
CL+Present	Head Noun	296	23
	Non-Head Noun	262	245
Physics+Describe	Head Noun	93	3
	Non-Head Noun	124	35
Physics+Present	Head Noun	185	3
	Non-Head Noun	192	93

TABLE 4-11. DIFFERENCE OF THE PROPORTIONS OF SCIENTIFIC-RESEARCH-GENERAL TERMS

Category	Chi-Square Value	Significant Level
CL+Describe	86.698	$p < 0.0001$
CL+Present	151.000	$p < 0.0001$
Physics+Describe	16.839	$p < 0.0001$
Physics+Present	67.449	$p < 0.0001$

TABLE 4-12. CHI-SQUARE STATISTICS

Category	Group	Frequency of Scientific-Research-General Term	Frequency of Research-Domain-Specific Term
CL	Head Noun	524	62
	Non-Head Noun	564	523
Physics	Head Noun	278	6
	Non-Head Noun	316	128

TABLE 4-13. COMPARISON BETWEEN PROPORTIONS OF SCIENTIFIC-RESEARCH-GENERAL TERMS

4.4.4 DISCUSSION

The result of the term distribution experiment clearly shows that in each of the four categories, with regard to the percentage of term frequencies taken by a certain proportion (5% and 10%) of terms, the head noun group is always larger than the non-head noun group. This is consistent with hypothesis I, which states that the head noun distribution should be narrower than the non-head nouns.

The results of the term generality experiment only show moderate agreement between human subjects. One possible explanation is that terms often have several different meanings. For example, the term ‘sense’ can both occur in papers of computational linguistics and of physics. However, we cannot say that ‘sense’ is a scientific research general term since it has different meanings in computational linguistic research domain and physics domain. The same applies to the words ‘translation’ and ‘frequency’. To solve this problem, instead of giving a single ambiguous word to the subjects, we can give the word and the context where it occurs (several neighbouring words). Another experiment we suggest as a future work would be to find whether a term is scientific-research-general by testing whether it occurs equally frequent in several different scientific research domains. Some preliminary work in this direction showed that this kind of test could only be based on very large corpora.

The comparison of the proportion of general terms between the head noun group and the non-head noun group strongly indicates that a head noun group tends to contain more general terms than a non-head noun group, as stated in hypothesis II.

These experiments focus on examining topic expressions only, while the relationships between different parts of a topic expression and the elements in the discourse are further explored in the next experiment.

4.4.5 EXPERIMENT ON NOUN PHRASES

The aim of this experiment is to verify whether in noun phrases, head nouns are more general than non-head nouns. To do so, we collect a set of noun phrases and split head nouns and non-head nouns into two groups. Like what we did above for topic expressions, we first compare the term distribution of the head noun group and the non-head noun group; we also verify whether the head noun group contain more general terms than the non-head noun group based on human judgement.

4.4.5.1 DATA COLLECTION

We use the same materials as described in section 4.4.2.1. These are documents or text snippets downloaded from the Institute of Physics and the ACL Anthology by searching cue phrases such as ‘this paper describes’ and ‘we present’. The raw materials are again grouped into four categories: ‘CL+Present’, ‘CL+Describe’, ‘Physics+Present’ and ‘Physics+Describe’. We use the NLProcessor to extract the noun phrases from the raw materials and remove those contained in topic expressions. Here topic expressions are identified in the same way as in section 4.4.2.1. The table below summarises the total number of noun phrases in each category.

Category	Number of Noun Phrases
CL+Describe	18270
CL+Present	3783
Physics+Describe	1861
Physics+Present	311

TABLE 4-14. NUMBER OF COLLECTED NOUN PHRASES

We use the same algorithm as described in section 4.4.2.1 to extract the head nouns and the non-head nouns from the noun phrases.

4.4.5.2 RESULT

Table 4-15 summarises the total number of unique terms and term frequencies in each group. The terms in each group are ordered by frequency of occurrences. Table 4-16 summarises the result of the term distribution analysis. Specifically, we want to examine the percentage of term frequencies taken by a certain proportion of top terms in a group (top 5% and top 10%). We see that it is consistent across all the categories that the percentage of the head noun group is higher than the non-head noun group. This indicates the head noun distribution is

Chapter 4 – Probing the Structure of Extended Topic

narrower than the non-head nouns and further suggests that head nouns are more general than non-head nouns.

Category	Group	Number of Terms	Term Frequency
CL+Describe	Head Noun	3943	18270
	Non-Head Noun	2066	6342
CL+Present	Head Noun	738	3783
	Non-Head Noun	401	1387
Physics+Describe	Head Noun	733	1861
	Non-Head Noun	352	836
Physics+Present	Head Noun	175	311
	Non-Head Noun	81	124

TABLE 4-15. NUMBER OF TERMS AND TERM FREQUENCIES IN NOUN PHRASES

Category	Group	5% Term	10% Term
CL+Describe	Head Noun	49.9%	62.9%
	Non-Head Noun	40.0%	52.9%
CL+Present	Head Noun	58.4%	68.1%
	Non-Head Noun	50.6%	58.7%
Physics+Describe	Head Noun	40.7%	50.5%
	Non-Head Noun	34.0%	46.9%
Physics+Present	Head Noun	24.8%	34.7%
	Non-Head Noun	17.7%	29.0%

TABLE 4-16. TERM DISTRIBUTIONS IN NOUN PHRASES

We also examine the proportion of scientific-research-general terms in different groups. To do so, we use the same set of human judged terms described in section 4.4.3.2. Table 4-17 summarises the number and the proportion of terms in each group that have human judgement result. Table 4-18 reports the frequency of scientific-research-general and research-domain-specific terms in each group. We see that the head noun group contains a larger proportion of scientific-research-general terms, which is true in each category. The differences are statistically significant as indicated by the chi-square values in Table 4-19.

Chapter 4 – Probing the Structure of Extended Topic

Category	Group	Number of Terms	Term Frequency	Proportion of Terms	Proportion of Term Frequencies
CL+Describe	Head Noun	143	4067	3.6%	22.3%
	Non-Head Noun	102	1187	4.9%	18.7%
CL+Present	Head Noun	83	821	11.2%	21.7%
	Non-Head Noun	61	450	15.2%	32.4%
Physics+Describe	Head Noun	65	347	8.9%	18.6%
	Non-Head Noun	35	112	9.9%	13.4%
Physics+Present	Head Noun	29	47	16.6%	15.1%
	Non-Head Noun	6	10	7.4%	8.1%

TABLE 4-17. STATISTICS OF THE REPRESENTATIVE TERMS IN NOUN PHRASES

Category	Group	Frequency of Scientific-Research-General Terms	Frequency of Research-Domain-Specific Terms
CL+Describe	Head Noun	2558	1509
	Non-Head Noun	350	837
CL+Present	Head Noun	557	264
	Non-Head Noun	148	302
Physics+Describe	Head Noun	269	78
	Non-Head Noun	53	59
Physics+Present	Head Noun	38	9
	Non-Head Noun	5	5

TABLE 4-18. DIFFERENCE OF THE PROPORTIONS OF SCIENTIFIC-RESEARCH-GENERAL TERMS IN NOUN PHRASES

Category	Chi-Square Value	Significant Level
CL+Describe	415.007	P<0.0001
CL+Present	143.787	P<0.0001
Physics+Describe	36.881	P<0.0001
Physics+Present	4.236	P=0.0396

TABLE 4-19. CHI-SQUARE STATISTICS FOR NOUN PHRASES

Based on the above experiments results, we conclude that it is generally true that in noun phrases, the head nouns are more general than the non-head nouns.

4.4.6 MAIN RESULT

The experiments described above have shown that there is a strong tendency for head nouns to be more general than non-head nouns in topic phrases. This result complies with the typical structure of extended topic, i.e., ‘the G of/that/for S’. We further show in section 4.4.5 that the above distinction between head nouns and non-head nouns is also observed in noun phrases in general. We are not aware of any linguistic theory that explains the distinction by comparing the functions played by head nouns and non-head nouns respectively. As one of the future direction, we can verify whether the theory of extended topic could be applied more broadly to noun phrases.

Term	Frequency
Method	120
Approach	90
Result	77
Model	53
System	50
Design	39
Experiment	27
Technique	25
Study	23
Development	18
Analysis	17
Feature	13
Overview	13
Application	12
Conclusion	11
Evaluation	10
Framework	10
Procedure	10
Solution	8
Investigation	7

TABLE 4-20. A LIST OF GENERAL TERMS AND THE FREQUENCIES IN BOTH ‘CL’ AND ‘PHYSICS’

The experiments help us to collect some typical generic concepts. We have collected the most frequent general terms in the head noun list in topic phrases, as shown in the table above. We

can see that many of these concepts such as ‘method’ and ‘result’ also occur in the question analysis section. Refer to appendix C.8 for more collected general terms.

4.5 EXPERIMENT III

4.5.1 INTRODUCTION

This experiment tests the idea that the generic part of an extended topic (i.e., a generic topic) has a strong relation with the new elements in the discourse, while the specific part of an extended topic (i.e., a specific topic) is closely related to the given elements in the discourse. To do so, we cannot just simply examine where (in the given elements or in the new elements) the terms in an extended topic occur in a relevant discourse. As stated in the last chapter, a generic topic is a general category indicating the kind of new information required in the discourse. For example, a generic topic ‘climate’ indicates that details about ‘temperature’ and ‘air pressure’ could be found in the discourse. Although the theory predicts that a specific topic directly corresponds to some given elements in the relevant discourse, however, many times it is not the specific topic itself that is repeated but some synonyms of it. For example, a discourse about the UK Conservative Party might use ‘the Tories’ to refer to it. It is expensive to model all these relations manually. We designed a method to automatically acquire conceptually related terms of a topic — i.e., its signatures (Hovy and Lin, 1998) — by observing term distributions. The concrete method is presented in section 4.5.2. Once the signatures are collected, we will show that the signatures of generic topics are more likely to refer to new information in the discourse than those of specific topics. The method for marking up the given entities and new entities in a discourse is also described in section 4.5.2.

4.5.2 MATERIAL AND METHOD

4.5.2.1 DOCUMENT PREPARATION

We collected 180 documents for 12 generic topics from the online medical references of University of Maryland²⁶, the Internet Drug List²⁷ and the Encyclopaedia Britannica 2004²⁸, with 15 documents for each generic topic. The general topics together with their related documents were divided into three groups. Several generic topics were grouped together

²⁶ Refer to <http://www.umm.edu/medref/index.html>

²⁷ Refer to <http://www.rxlist.com/>

²⁸ Refer to <http://www.britannica.com/>

Chapter 4 – Probing the Structure of Extended Topic

because they are associated with the same kind of specific topic. The table below gives all the generic topics and the kind of associated specific topics.

	Generic Topics	Associated Specific Topic
Group 1	Cause, prevention, symptom, treatment	A disease
Group 2	Pharmacology, dosage, precaution, side effect	A drug
Group 3	Climate, cultural background, ethnic composition, economic resources	A country

TABLE 4-21. GROUPS OF GENERIC TOPICS AND ASSOCIATED SPECIFIC TOPICS

Documents of different generic topics in the same group were also aligned according to their specific topics. For example, 'acne' is one specific topic in group 1, so there is a document about 'the causes of acne', a document about 'the prevention of acne', a document about 'the symptoms of acne' and a document about 'the treatment of acne'. Refer to table 4-22 for more examples. Each column in table 4-22 corresponds to one specific topic and each row corresponds to one generic topic. To sum up, each generic topic indexes 15 documents, all of which differ in their specific topics; each specific topic indexes 4 documents, all of which differ in their generic topics. Example documents of each group are included in appendix D.1.

	Acne	Adolescent depression	Acute mountain sickness	...
Prevention	The prevention of acne	The prevention of adolescent depression	The prevention of acute mountain sickness	...
Symptom	The symptoms of acne	The symptoms of adolescent depression	The symptoms of acute mountain sickness	...
Cause	The cause of acne	The cause of adolescent depression	The cause of acute mountain sickness	...
Treatment	The treatment of acne	The treatment of adolescent depression	The treatment of acute mountain sickness	...

TABLE 4-22. DOCUMENTS IN GROUP1 AND THEIR TOPICS

4.5.2.2 SIGNATURE SELECTION

In section 4.5.1, we introduce the need for automatically constructing topic signatures by observing term distributions. Previous studies (Hovy and Lin, 1998) fulfil this task by choosing the most frequent terms in the documents on topic (a large set of documents is used to tolerate individual phenomena). In our case, we only collected a small number of documents

for each topic and each document has two topics (i.e., the generic topic and a specific topic), so the fact that a term is frequent in the document set could be due to the shared topic, but could also be due to the special topic of an individual document. Therefore, instead of choosing the most frequent terms, we chose the most evenly distributed terms from these that meet a low coverage threshold. Our topic signature selection procedure can be summarised in the following steps:

- a. examine all the documents about one topic (i.e., 15 documents for each generic topic and 4 documents for each specific topic);
- b. remove less content-bearing terms based on POS tags, so the remaining terms include nouns, verbs, participles, adjectives, numbers, adverbs and prepositions;
- c. select words that meet a low coverage threshold (i.e., occur in at least 3 documents for generic topics and at least 2 documents for specific topics);
- d. choose the most evenly distributed terms.

Below are the formulas for calculating how evenly a term is distributed across several documents, in which c_i means the number of times for a term to occur in one document.

$$c_{i-smooth} = c_i + 0.2 \quad (4.1)$$

$$c_{i-norm} = c_{i-smooth} * 100 / \sum_i c_{i-smooth} \quad (4.2)$$

$$Evenness = \frac{1}{\sqrt{\sum_i c_{i-norm} * c_{i-norm}}} \quad (4.3)$$

Through this procedure, we chose about 30 words for each generic topic and about 15 words for each specific topic²⁹. For example, the signatures of a generic topic ‘side effects’ and of a specific topic ‘methadone hydrochloride’ are shown as follows. The signatures of other topics are included in appendix D.2.

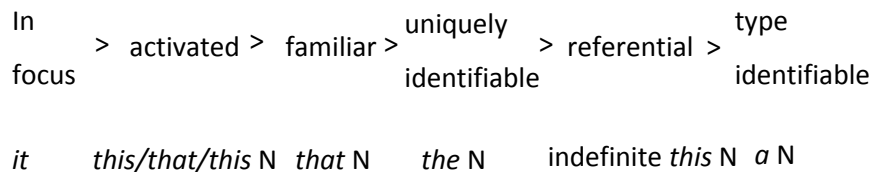
²⁹ The reason why choosing more signatures for generic concepts is to make sure that the total number of occurrences of the generic topic signatures in the document collection and that of the specific topic signatures are similar. The numbers 30 and 15 are not the exact numbers, because the evenness score does not provide an absolute ordered list to allow us to cut at any point we want.

Signatures ('side effect')=[headache, adverse, diarrhea, cardiovascular, drug, nausea, system, reaction, effect, libido, hydrochloride, side, following, much, use, mouth, insomnia, dizziness, constipation, rash, edema, gastrointestinal, nervous, urticaria, hypotension]

Signatures ('methadone hydrochloride')=[Prominent, central, administer, slow, syndrome, maintenance, when, analgesic, sedation, methadone, severe, narcotic, hydrochloride, nervous]

4.5.2.3 GIVENNESS ANNOTATION

Brown and Yule (1983: 169-179), Prince (1992) and Gundel (2003) discuss the linguistic forms that signal given and new elements. Brown and Yule (1983: 171) find “two predominant forms of expression used to refer to an entity treated as given, pronominals and definite NPs”. Prince (1992) notes that definite articles, demonstratives, possessives, personal pronouns, proper nouns, quantifiers like ‘all’, ‘every’ signal definiteness and indefinite articles, quantifiers like ‘some’, ‘any’ and ‘one’ signal indefiniteness. Gundel (2003) provides a givenness hierarchy as below.



Based on these theories, we designed the following heuristics for marking up the given and the new elements. Documents were pre-processed using Connexor’s syntax analyser³⁰ to acquire all the required syntactic and morphological information.

- a. terms that follow definite articles, demonstratives and possessives are GIVEN;
- b. terms that are quantified by ‘all’ and ‘every’ are GIVEN;
- c. terms that are determined by indefinite articles are NEW;
- d. numbers and number quantified terms are NEW;
- e. terms that follow ‘some’, ‘any’, ‘one’ and ‘a few’ are NEW;

³⁰ Refer to <http://www.connexor.com>

Chapter 4 – Probing the Structure of Extended Topic

f. terms that have been mentioned in the title or the foregoing discourse is GIVEN;

g. all other terms are marked as BLANK.

The above heuristics were applied according to the priority: a, b, c, d, e > f > g. Therefore, if a term is inferred as referring to a new entity according to either c, d or e, even it has occurred before in the forgoing discourse (i.e., meet f), the current occurrence is still marked as new. The above algorithm marks the given/new distinction at the sentence level. From the discourse point of view, a new element might first be introduced as a new element but later on be referred as a given element. Brown and Yule (1983: 169) note that “it has been observed that, in English, new information is characteristically introduced by indefinite expressions and subsequently referred to by definite expressions.” Therefore, to decide whether a term introduces a new element from the discourse point of view, we only considered its first occurrence, except if it refers to another new entity in the succeeding occurrences.

4.5.3 RESULT

The numbers of occurrences of topic signatures in different categories are shown in the tables below.

	Treatment	Symptom	Prevention	Cause	Generic Topic	Specific Topic	Total
GIVEN	27	19	16	45	107	162	269
NEW	19	2	9	23	53	42	95
BLANK	168	83	102	131	484	275	759
NEW+BLANK	187	85	111	154	537	317	854

TABLE 4-23. THE NUMBERS OF OCCURRENCES OF TOPIC SIGNATURES IN DIFFERENT CATEGORIES—GROUP 1

	Side effect	Precaution	Pharmacology	Indication	Generic Topic	Specific Topic	Total
GIVEN	39	26	31	42	138	137	275
NEW	4	20	12	48	84	69	153
BLANK	124	135	141	164	564	351	915
NEW+BLANK	128	155	153	212	648	420	1068

TABLE 4-24. THE NUMBERS OF OCCURRENCES OF TOPIC SIGNATURES IN DIFFERENT CATEGORIES—GROUP 2

Chapter 4 – Probing the Structure of Extended Topic

	Climate	Ethnic Composition	Cultural Background	Resources	Generic Topic	Specific Topic	Total
GIVEN	52	134	59	51	296	189	485
NEW	210	77	29	10	326	23	349
BLANK	138	110	163	189	600	323	923
NEW+BLANK	348	187	192	199	926	346	1272

TABLE 4-25. THE NUMBERS OF OCCURRENCES OF TOPIC SIGNATURES IN DIFFERENT CATEGORIES—GROUP 3

The above tables show that compared to specific topics, a larger proportion of generic topic signatures introduce new entities. The chi-square statistics suggest that such differences are significant for table 4-23 [chisquare=7.307, $p < 0.01$] and table 4-25 [chisquare=112.235, $p < 0.0001$], but not so for table 4-24 [chisquare=0.877, $p > 0.05$]. A problem here is that there are a large number of BLANK elements left unmarked. In our algorithm, there are only a few linguistic forms that reliably signal new elements. Therefore, it is reasonable to consider most terms annotated as BLANK are actually new elements. We added up the blank terms and the new terms, as shown in the fourth row of each table. In this case, the specific topic signatures are significantly more probable to be a given element in the discourse than the generic topic signatures [Table 23: chisquare=44.640, $p < 0.0001$; Table 24: chisquare=9.918, $p < 0.001$; Table 25: chisquare=22.960, $p < 0.0001$].

4.5.4 DISCUSSION

The result of the experiment indicates that, in general, the generic topic of a discourse has stronger relations with the new elements in the discourse than the specific topic. However, we can see that the result is not consistent across different generic topics. For example, the signatures of generic topic ‘side effect’ refer to 39 given entities but only introduce 4 new entities. The ratio between the given and new entities is much higher than other generic topics and is even higher than average ratio of the specific topics in the same group.

One major drawback of the experiment lies in the given/new annotation step. As observed in section 4.5.3, there are still a lot of elements that our algorithm cannot determine.

Another problem lies in the signature selection method. Since there are only a small number of documents for each topic, many terms are selected by coincidence rather than a strong relation with the topic. However, there is probably no easy method to solve this problem.

4.6 SUMMARY

In section 4.2, we collect a corpus of WH-questions and recast them into the form of “what is/are the G of/that/for S”. We extract the concepts in the generic part and classify them into different types based on their complexity and how they fulfill the function of navigating in a knowledgebase. We suggested that those concepts that retrieve a network of related events or propositions play an important role in discourse planning and information retrieval.

In section 4.3, we illustrate that given a generic concept, human subjects are able to tell relevant passages from irrelevant ones. This shows that generic concepts could function as a knowledge retrieving facility in human brain.

The experiment in section 4.4 is based on a collected set of topic expressions describing the plan of academic papers. It shows that the head nouns of such topic expressions contain a larger proportion of general term than the non-head nouns. The result suggests that the typical structure of extended topic, that is, “the X of/that/for Y” also underlies such topic expressions.

The experiment in section 4.5 focuses on investigating the relations between different parts of topic expressions and different discourse constituencies. It shows that compared to specific topic, the signatures of generic topics contain a larger proportion of terms referring to new elements in the discourse, although for some particular generic concepts, the data is insufficient to support this conclusion.

The results of all the experiments provide empirical supports for the theory of extended topic.

CHAPTER 5

APPLICATIONS OF THE THEORY OF EXTENDED TOPIC

5.1 INTRODUCTION

In previous chapters we developed a theory of extended topic and designed a series of empirical studies to validate the theory. In this chapter we will explore the potential applications of the theory. Several research areas will be introduced in turn in section 5.2 to 5.6, including knowledge indexing, discourse segmentation, discourse planning, generating indicative topic expressions and information retrieval. Section 5.7 provides a summary of the chapter.

5.2 KNOWLEDGE AND TEXT INDEXING

Knowledge and text indexing aim to facilitate the retrieval of relevant information by building effective indices to knowledge or textual content. There are two typical approaches to index building: one is to apply a pre-defined meta-schema; the other is to directly select key words (or key phrases) from texts based on term or phrase weights. In the second approach, the weight of a term or a phrase is usually calculated based on its frequency of occurrences; detailed method will be introduced in Chapter 6. The meta-schema in the first approach is typically represented as a concept hierarchy. For example, Lykke and Eslau (2010) use a thesaurus to index medical domain documents, which is a hierarchical structure and terms inside are linked by *broader-term* or *narrower-term* relationship. Appel et al. (1988) define taxonomy to represent objects and relationships in medicine which is applied to indexing biomedical literature. In both of the above work, the index is typically a single concept that does not have a structure as extended topic. We suggest using extended topic to define text or knowledge indices. A few reasons are elaborated below.

First, extended topics could match to user queries. As discussed in section 3.2, the notion of extended topic is formalised by probing the characteristics of user queries to a knowledge base. In section 4.2, we also show that the structure of extended topic could be observed in many WH-questions. Here we would highlight two attributes. To match to user queries, a text index must not contain any information new to the user. Extended topics meet this requirement since the specific part of an extended topic corresponds to the given information in the discourse and the generic part is too general to convey anything new. One could

certainly argue that, for the same piece of information, different users may take different parts as given and new. For example, for the same piece of information, ‘Mary was studying biology in Duke University’, we can formulate different questions such as ‘Which university did Mary study biology in?’ and ‘What discipline did Mary study in Duke University?’. Thus, to meet different user needs, we should include all the different indices generated by presuming different parts as new information. Another attribute is that a good text index should characterise the kind of new information that users may want. Most text indexing methods simply select prominent terms based on shallow features such as term frequencies and positional information, which may be effective in generating focused entities (i.e., the specific part of extended topic) but would leave out the kind of required information about the focused entities (i.e., the generic part of extended topic).

Secondly, the structure of extended topic avoids ambiguities that exist in many topic expressions containing a general concept only. For example, if we mark a discourse as being about ‘biography’, then users would not be able to tell whether it contains biographical facts of a particular person (e.g., his education and career development) or it talks about some general attributes of biography as a type of written work (e.g., key facts in a typical biography). However, the ambiguity would be resolved if the concept ‘biography’ is used under the context of an extended topic. Specifically, when biography is the generic part of an extended topic such as in ‘the biography of Charles Dickens’, we would anticipate the discourse to list biographical facts of a person; in contrast, when biography is the specific part of a topic expression such as in ‘skills in writing good biographies’, we would anticipate the discourse to talk about the general attributes of a biography as a written work. As mentioned in section 3.2, generic topic and specific topic are two roles; given a general concept such as ‘biography’, users would have different prediction of the discourse content depending on which role they presume the concept plays. In appendix E, we provide example passages to illustrate the difference between different topic expressions.

5.3 TEXT SEGMENTATION

As discussed in section 2.3.1, in some linguistic studies, such as van Dijk (1977), the notion of ‘discourse topic’ is defined as a discourse organisation principle. Van Dijk (1977) notes that the facts denoted by sentences are inter-related in virtue of some “COMMON BASIS” known as the discourse topic. Further, he represents the macrostructure of a discourse by a tree with each node denoting a discourse topic that organises a text segment. Although the structure of

extended topics is quite different from discourse topics as defined in van Dijk (1977), we will provide a few sample discourse plans and short passages to show that extended topics are also used to structure the discourse.

The discourse organisation principle functions on two levels: the macro level and the micro level. The macro level is concerned about the high-level structure of the whole discourse. In other words, it addresses how sections or passages relate to each other. At the macro level, a discourse about a specific topic can often be segmented into several parts, with each part focusing on a certain perspective (a generic topic) of the specific topic. In section 3.3.3, we show that extended topics often appear in section headings, sentences describing the plan of a document, entries in the index of a book and the table of content of an article. For example, below is the table of contents of an article about Seattle shared on Wikipedia³¹. We see that the whole article is about a specific topic (Seattle) and that each section or subsection is about one perspective of Seattle, such as its history or geography.

[5.1] Contents

- 1 History
 - 1.1 Founding
 - 1.2 Timber town
 - 1.3 Gold Rush, World War I, and the Great Depression
 - 1.4 Post-war years: aircraft and software
- 2 Geography
 - 2.1 Topography
 - 2.2 Surrounding municipalities
 - 2.3 Climate
 - 2.4 Neighborhoods
- 3 Cityscape
 - 3.1 Landmarks
- 4 Culture
 - 4.1 Performing arts
 - 4.2 Media
 - 4.3 Tourism
 - 4.4 Sports

³¹ Cited from <http://en.wikipedia.org/wiki/Seattle>

- 4.5 Outdoor activities
- 5 Economy
- 6 Demographics
- 7 Government and politics
- 8 Education
- 9 Infrastructure
 - 9.1 Health systems
 - 9.2 Transportation
 - 9.3 Utilities

The micro level concerns the coherence within a section or a passage. One important determinant of the coherence of a passage is whether neighbouring sentences deal with one perspective of a specific topic. For example, in Passage [5.2] (a sample text extracted from Hahn (1990)), the passage appears coherent since sentences b, c and d all concern the configuration of the Delta-X system. In contrast passage [5.3], generated simply by changing the order of the propositions in passage [5.2], appears less coherent because sentence c no longer talks about configuration.

[5.2] a. The Delta-X is a computer system from ZetaMachines Inc. b. The system is based on a 68020 processor. c. It has a 12-inch monochrome display and an integrated telephone handset and built-in modem. d. Internally, there is a 40-megabyte hard disk, a 1.2-megabyte 5.25-inch floppy disk drive, 4.5 megabytes of RAM, three RS-232C ports, and an ST-506 port....

[5.3] a. The Delta-X is a computer system based on a 68020 processor. b. It has a 12-inch monochrome display and an integrated telephone handset and built-in modem. c. It is from ZetaMachines Inc. d. Internally, there is a 40-megabyte hard disk, a 1.2-megabyte 5.25-inch floppy disk drive, 4.5 megabytes of RAM, three RS-232C ports, and an ST-506 port....

Many text segmentation methods are based on cohesion theory (Hearst, 1997; Kan et al., 1998; Morris and Hirst, 1991). These methods calculate the linkage between textual units based on word repetitions, word co-occurrences and referential links, etc. Such methods focus on the shift in specific topic but omit the role of generic topic. Many generic concepts denote a family of relationships. For example, the concept anatomy denotes a set of compositional

and spatial relationships. A text segmentation method could model these relationships to detect shifts in generic topic.

5.4 DISCOURSE PLANNING

The discourse planning phase in a text generation system is concerned with “imposing ordering and structure over the set of messages to be conveyed” (Reiter and Dale, 1997). A typical discourse plan is a tree structure: the leaf nodes of the tree represent a message to be included in the discourse and the internal nodes represent the relationships between the messages. Previous approaches for discourse planning could be divided into two types: the planning-based approach and the schema-based approach. Young and Moore (1994) is one representative work of the planning-based approach. The discourse plan contains an ultimate communication goal on the top level of the tree and a list of actions on the second level which would fulfill the communication goal together. Each action would further be decomposed into a list of sub-actions at a lower level. In the preparation stage, the system has a set of pre-defined actions, each of which is associated with the preconditions to take the action, the effects of the action and a decomposition plan. When discourse planning starts, the system dynamically select actions to meet the overall communication goal. There are a few factors that influence the action selection decisions, including whether the effect of an action could help achieve the communication goal, whether the pre-conditions of an action are met, and the discourse length to determine whether an action should be further expanded to smaller actions. The planning-based approach is domain independent; the schema-based approach, rather, would only fit for a small domain. A schema is a pre-defined text plan fitting for a stereotypical scenario or communication purpose. It is composed by a sequence of messages or smaller schemas. In the discourse planning stage, the system just needs to pick up which schema to apply. One representative work of this approach is by Mckeown (1985), who presents a system to generate answers to user questions. They define schemata to fulfill typical question types by rhetorical strategies such as identification and attributive. For example, question ‘What is a ship?’ is mapped to the identification schema.

The discourse plan in Young and Moore (1994) is driven by a given communication goal, which could be exemplified as a proposition that the speaker wants the hearer to believe. Mckeown (1985) focuses on typical questions types and define rhetorical strategies to organise answers to different questions. In comparison, instead of having one specific goal to achieve, a long narrative or expository prose simply aims to share a lot of detailed knowledge about a

particular topic. Example 5.1 provides the structure of an article about Seattle. It is clear in this example that topic plays a key role in discourse organisation; specifically, we see that each section is about one particular perspective about Seattle, including Seattle history, Seattle culture, etc. In section 5.3, we suggest that extended topics help organise a discourse at both a macro level and a micro level. Thus we propose to apply the notion of extended topic to help discourse planning in the following way. At the macro level, a general discourse structure schema could require each section to focus on a particular perspective (generic topic) of a specific topic. At the micro level, the content of a passage would be chosen by retrieving knowledge about the perspective from the knowledge database. This approach assumes that knowledge stored in the database or knowledge base is already indexed by extended topic. Kan et al. (2001) use topic trees to index the content of medical domain documents and to plan the structure of cross-document summaries. Figure 5-1 is an example topic tree that they provide. We see that the second layer of the tree is one particular heart disease ‘angina’ and the third layer contains different perspectives of angina such as causes and treatment. The second layer and the third layer together would formulate extended topics such as ‘the causes of angina’ and ‘the treatment of angina’. To generate summary for multiple documents, their system would first compare the topic trees of the documents to identify shared topics. The generated summary is structured as a list of shared topics, each of which is then replaced by sentences (about the topic) selected from the original documents.

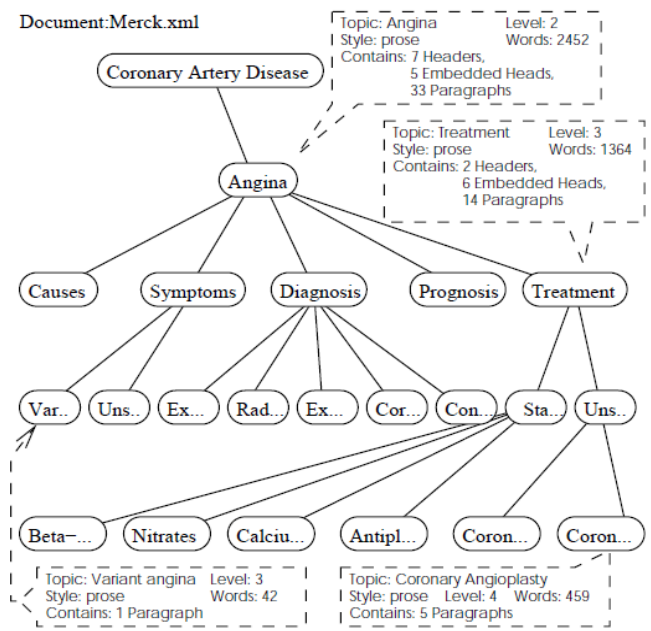


FIGURE 5-1. A TOPIC TREE FOR AN ARTICLE ABOUT CORONARY ARTERY DISEASE (KAN ET AL., 2001)

5.5 GENERATING INDICATIVE TOPIC EXPRESSIONS FOR PASSAGES

The theory of extended topic formalises the structure of indicative topic expressions. Most current topic generation systems, as mentioned in section 2.3.2, focus on the task of extracting the most *salient* elements from a document. They typically use shallow features to detect *salience*, such as position within a sentence, term frequencies, cue phrases, and strength of lexical chains. These methods could effectively extract the specific topic but not the perspectives by which they are approached. We suggest the following steps to generate extended topic expressions. Given a list of potential perspectives (the generic concepts), each associated with a characteristic set of relations and perspective patterns, a program could generate a topic as follows: first, identify the centered entity (e.g., by frequency of mention); second, find the perspective most strongly evidenced by the discourse connectives, clause patterns, formatting, etc.; finally, combine the perspective and the centered entity and generate a title phrase. The same method could be employed to create a discourse-deictic reference to the passage, or an entry for an analytical index.

5.6 AUTOMATIC RETRIEVING INFORMATION FROM DOCUMENTS

Typical information retrieval (IR) systems are based on simple key word matching. They treat the key words in the generic part and the specific part of a query equally. The theory of extended topic suggests that the generic part of an information request will be replaced in the answer by information that is *new* to the questioner, while the specific part will be kept as a *known* part. Therefore, to match a document to an information request, we should use the *given* part (i.e. the *known* part) in the document to match the specific part of the information request and use the *new* information to match the generic part. This idea, which relates the generic/specific distinction in an information request to the given/new distinction in a discourse, has never been applied before in research on IR. Below we modify the language model IR approach to incorporate the mapping relationship. The language model approach scores a document based on $P(q_i|D)$, i.e., the probability for a document D to generate a query term q_i . We could re-define the formula to have two parts: $P(s_i|Given)$ and $P(g_j|New)$. $P(s_i|Given)$ models the probability for the given information in a document to generate a query term s_i in the specific part of an information request; $P(g_j|New)$ models the probability for the new information in a document to generate a query term g_j in the generic part of an information request. $P(s_i|Given)$ is easy to calculate since, based on the theory of extended topic, the specific part of an information request directly maps to the given information in a

relevant document. However, it would be much harder to match between the generic part and the new information. A simple approach is the one introduced in section 4.5, which is applied in the experiments on verifying the mapping relationship between different components of an extended topic and different components in a document. This approach analyses word co-occurrence patterns to extract a list of signature words for a generic concept, which could then be matched to the new information in a document. Another problem is how to separate the *given* and the *new* information in a document. Again, we could apply the approach introduced in section 4.5.2.3, which uses simple lexical or syntactic cues to achieve the above purpose.

Above we talk about adapting the typical word frequency based IR approach to incorporate the mapping relationships between different parts of an extended topic and different constituencies in a document. A key problem is to match the generic part in the extended topic to the new information in the document. We propose a simple solution which extends the generic part by a list of signature words. This solution maintains the key-word-based IR retrieval model but has a limitation in modelling complex relationships. Below we will further discuss this problem to seek more refined solutions.

In section 4.2, we extract a set of generic concepts from WH-questions and classify them into different classes. Some general concepts represent a simple attribute, such as price and colour. If the instances of such an attribute are enumerable, the mapping relationships between the attribute and the instances could be easily encoded in an ontology or a thesaurus. For example, the relationship between the concept 'colour' and particular colours such as green and yellow could be modelled in a thesaurus. For attributes that are not enumerable, patterns defined by lexical and syntactic features could be applied to extract such attributes. For example, '<NUM> \$' could be used to mark all the prices. Some generic concepts retrieve an entity/event/proposition by its relationship with other entity/event/propositions or prime a family of relationships; we suggest the following steps to sharpen the retrieval of such generic concepts.

1. Choose the desired perspective (e.g., *procedure*, as in 'procedure for making an omelette').
2. Identify a set of relevant relations for the chosen perspectives (e.g., purpose, steps, precondition, instrument);
3. Identify *perspective patterns* through which these relations are expressed in documents, including discourse connectives, types of clause, and formatting. For procedural relations

these would include imperatives, the discourse connectives found in procedural programming languages (if, when, until, etc), forms like ‘to do <action>’, ‘by doing <action>’, ‘use <instrument>’, ‘with <instrument>’, and enumerated and bulleted lists.

4. Use conventional IR methods based on key words in order to retrieve passages containing the focused entity.
5. Use perspective patterns in order to find and analyse passages in which the focused entity is approached from the desired perspective.
6. Return the relevant information, either by listing the selected passages, or (if possible) by integrating them.

5.7 SUMMARY

In this chapter we discuss how the theory of extended topic could be applied to various research areas. In the rest of the thesis we will focus on applying the theory to facilitate information retrieval. Above we propose that the typical word-frequency-based IR approach may be better suited to retrieving the specific part of an extended topic than the generic part. One key problem lies in modelling the relationship between the generic part in the information request and the new information in the discourse. For generic concepts that correspond to a relationship or family of relationships we propose, as a general method, to identify perspective patterns to detect the relationship(s). In the next chapter, we will introduce the literature on IR, elaborate on why current IR models are ill-suited to retrieving the generic parts of extended topics, and propose detailed solutions. In chapter 7 and 8, the proposed solutions will then be applied in experiments on retrieving several types of extended topic.

CHAPTER 6

APPLY EXTENDED TOPIC TO INFORMATION RETRIEVAL

6.1 INTRODUCTION

In chapter 2 to 5, we establish the theory of extended topic and also discuss how it could be applied to help a few application areas. In particular, in the application area of information retrieval (IR), we point out that the typical key word matching based IR approach may be better fitted to retrieve documents relevant to the specific part than the generic part of an extended topic. In chapter 5, we propose a general methodology to improve retrieving the generic part.

In this chapter, we will discuss in details about applying the theory of extended topic to improve information retrieval. We will first introduce various information retrieval models; then analyse why they fit better for retrieving the specific part than for retrieving the generic part. As pointed out before, one problem is that the generic part is likely to be a hidden concept in the document. Direct key word matching cannot find the match. The query expansion technique is typically applied in information retrieval systems to expand the original query to include related concepts and therefore could partly solve the aforementioned problem. We introduce typical query expansion methods and also discuss their limitations. In the middle of this chapter, we expand the general methodology to improve retrieving the generic part described in chapter 5 and propose different methods to address different types of extended topic (as defined in chapter 4). The methods proposed are similar, in different respects, to several existing research areas, such as phrasal retrieval, text categorisation (TC) and question answering (QA). We also briefly introduce the literatures in these areas.

6.2 BASIC IR MODELS

6.2.1 INTRODUCTION

Information retrieval (IR) is the core technology for locating relevant information in online texts and digitised documents. An IR system makes indices to represent the documents. To retrieve relevant documents in the search space, the system calculates the similarity or the relevance between the query and the indices of the documents. So the indices of a document form a “*logical view*” of it (Baeza-Yates and Ribeiro-Neto, 1999: 5-15, 191-228). Different

types of indexing items have been explored, from strings and key words to concepts, and further to semantic constructs generated by applying NLP techniques to process documents. However, linguistic analysis is expensive to implement and there is no empirical proof showing that it could improve IR. Currently, the key-word-based indexing technique is still the most commonly adopted approach in IR. Thus, instead of applying deep syntactic or semantic analysis to process documents, a majority of work was done by simply exploiting words and word frequencies. It is evident that this simplification reduces the semantics of a document to “lexical semantics” and disregards the role played by “compositional semantics” (Sebastiani, 1999). Nonetheless, key-word-based retrieval is proven to be cheap and effective. The basic assumption underlying this approach is that word distributions are reflective of the hidden document topics. So the question is how to exploit the word distribution information to well represent the topics of documents. This section will introduce a few classical IR models, including the Boolean model, the vector space model, the Bayesian network model, the probabilistic model and the language model.

6.2.2 BOOLEAN MODEL

The Boolean model is the first IR model. According to Hiemstra (2001), it was “the leading model for commercial retrieval systems until the mid 1990’s”. The Boolean model is based on the set theory and the Boolean algebra. The query expressions formulated by users define a set of documents. For example, term ‘biology’ defines the set of documents that contain term ‘biology’. Query terms could be connected by three Boolean operators to form complex queries, including *not*, *and* and *or*. The figure below lists a few example queries and the sets that they define.

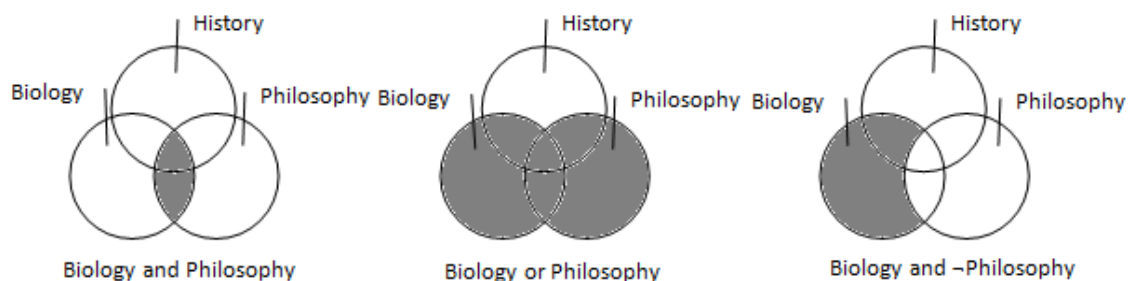


FIGURE 6-1. BOOLEAN QUERY EXPRESSION EXAMPLES

In the above figures, a circle represents a set of documents defined by a term. There are eight non-overlapping regions defined by the three circles. The union of the eight regions would give 256 combinations of the different regions and thus retrieve 256 sets of documents.

The main disadvantage of the Boolean model is that it does not output a ranked list of documents; instead, it sets a binary decision on whether a document is relevant. Salton et al. (1983) introduce the *extended Boolean model* which solves this problem. Details are described below.

First, for queries that contain one single term, instead of simply marking the set of documents that contain the term as relevant, the extended model will assign a score between 0 and 1 to a document by taking the term frequency into consideration. One way is to use the normalised tf-idf schema, where tf refers to the number of occurrences of a term x in a document and idf refers to the inverse document frequency of the term x in the document collection, as shown in formula 6.1. To ensure that the term score falls between 0 and 1, the original tf-idf formula is typically normalised by the maximum tf and idf value, as shown in formula 6.2.

$$idf(x) = \log(N / n(x)) \quad (6.1)$$

$$tf(x) * idf(x) = (tf(x) / \max_l(tf(l))) * (idf(x) / \max_i(idf(i))) \quad (6.2)$$

In formula 6.1, $n(x)$ refers to the number of documents that contain term x and N refers to the total number of documents in the collection. In formula 6.2, l refers to a term in a document and $\max_l(tf(l))$ refers to the frequency of the most frequent term in the document; i refers to a term in the document collection and $\max_i(idf(i))$ refers to the maximum idf value of all the terms in the document collection.

Above we derive a score for each term and document pair; further for two terms x and y , we can define a two dimension space and a document can be mapped to this space by (x, y) . The logical *or* operator of the two terms scores the documents based on the distance between the document and the point $(0, 0)$ (defined in formula 6.3) and the logical *and* operator scores the document based on the complement of its distance to point $(1,1)$ (defined in formula 6.4). Both 6.3 and 6.4 add a factor $1/\sqrt{2}$ to normalise the initial distance value to be below 1. The above definitions use the 2-norm Euclidean distance function. The model could be extended to n-dimensional using Euclidean distance functions, as defined in formula 6.5 and 6.6. When a query contains both *or* and *and* operators, such as $q = (x_1 \text{ and } x_2) \text{ or } x_3$, the Euclidean distance is defined as in formula 6.7.

$$Score(q_{or}, d) = \sqrt{\frac{x^2 + y^2}{2}} \quad (6.3)$$

$$Score(q_{and}, d) = 1 - \sqrt{\frac{(1-x)^2 + (1-y)^2}{2}} \quad (6.4)$$

$$Score(q_{or}, d) = \sqrt{\frac{x_1^2 + x_2^2 + \dots + x_n^2}{n}} \quad (6.5)$$

$$Score(q_{and}, d) = 1 - \sqrt{\frac{(1-x_1)^2 + (1-x_2)^2 + \dots + (1-x_n)^2}{n}} \quad (6.6)$$

$$Score(q, d) = \left(\frac{\left(1 - \left(\frac{(1-x_1)^2 + (1-x_2)^2}{2} \right)^{1/2} \right)^2 + x_3^2}{2} \right)^{1/2} \quad (6.7)$$

We see that the extended Boolean model assigns a score to each document, based on which documents could be ordered into a ranked list; thus, it overcomes the main disadvantage of the Boolean model, which only outputs a set of relevant documents.

6.2.3 VECTOR SPACE MODEL

The vector space model scores a document based on its similarity to the query. As its name indicates, the vector space model represents all documents and queries as vectors in a high-dimensional space. The similarity between two vectors is defined as the vector distance. The first challenge is to define a set of orthogonal (linear independent) dimensions. We could first think of defining dimensions by extracting a list of core concepts (or topics) from the corpus. However, it is not easy (if not unachievable) to define a perfect concept (or topic) list. In practice, most systems, such as in Salton and McGill (1983), just extract all the key terms or key phrases in the corpus, each of which corresponds to one dimension. This approach clearly ignores the distributional and semantic relationships among terms.

There are different approaches to calculate the value of a vector at each dimension. The simplest way is to assign binary value based on whether a term is present in a document. This

approach ignores multiple occurrences of a term in a document. Another approach is to use tf, i.e., the frequency of a term in a document. This approach does not consider the fact that different terms vary in their importance. Therefore, a term weighting schema is often plugged in to refine the raw value of term frequencies.

TERM WEIGHTING SCHEMA

The best known term weighting schema is the tf-idf schema, as shown in formula 6.2. The raw values of tf and idf are often normalised by the maximum value or transformed by a monotone increasing function to ensure overall document score is not influenced too much by one single term, one of such implementation is as shown in formula 6.2. In addition, document length is often taken into consideration as the third factor. Specifically, the original weighting schema will multiply a decreasing function of the document length to avoid long documents being favoured over short documents.

An alternative approach is to develop a model for approximating the word distribution and use this model to characterise its importance for retrieval. That is, we wish to estimate $P_i(k)$, the chance or the proportion of times that word w_i appears k times in a document. Bookstein and Swanson (1974) and Harter (1975a) study the method to select good indexing words. They observed that “function” word tends follow the Poisson model while a “content-bearing” word does not. Presumably for a function word, each token is allocated randomly to a document in a document collection following the Bernoulli process³²; when the number of documents gets bigger, the document distribution could be approximated by the Poisson distribution³³. Therefore, we could model the probability for a function word w to appear k times in a document as $p(k; \lambda)$, where λ is the average number of occurrences of w per document. Moreover, they present that a content-bearing word typically follows a 2-Poisson distribution. Specifically, there are two sets of documents for a content-bearing word: the *elite* set with a high average number of occurrences and the *non-elite* set with a low average number of occurrences. The distribution of the term in each set follows the Poisson distribution, thus derives a 2-Poisson model, as shown in the formula below.

³² Refer to http://en.wikipedia.org/wiki/Bernoulli_process

³³ Refer to http://en.wikipedia.org/wiki/Poisson_distribution

$$p(k; \pi, \lambda_1, \lambda_2) = \pi p_{elite}(k, \lambda_1) + (1 - \pi) p_{nonelite}(k, \lambda_2) \quad (6.8)$$

Where π is the probability of a document being in the elite set; $1 - \pi$ is the probability of a document being in the non-elite set; λ_1 and λ_2 are the average numbers of occurrences of word w in the elite set and non-elite set respectively.

Harter (1975a) applies the above model to calculate the effectiveness for a term to be a good indexing word in the context of document retrieval, as shown in formula 6.9. The underlying presumption is that the bigger the difference between the elite set and the non-elite set, the more likely for a term to be a content word. Van Rijsbergen (1979) scores a document using the probability for the document to be in the elite set given the term frequency in the document, as indicated in formula 6.10 for term weighting.

$$\frac{\lambda_1 - \lambda_2}{\sqrt{\lambda_1 + \lambda_2}} \quad (6.9)$$

$$\frac{\pi p_{elite}(k, \lambda_1)}{\pi p_{elite}(k, \lambda_1) + (1 - \pi) p_{nonelite}(k, \lambda_2)} \quad (6.10)$$

The biggest problem for applying the 2-Poisson model is the need to estimate three parameters for each term. Here we will not elaborate on the detailed methodology to estimate the parameters.

VECTOR DISTANCE CALCULATION

As noted above, the vector space model scores documents based on vector distances. There are various schemata to calculate vector distances, as shown in the below table, where q_i stands for the i th entry of the query vector and d_i stands for the i th entry of the document vector.

The vector space model is widely adopted in IR experiments. For example, the SMART system (Salton and McGill, 1983) was first developed at Harvard in the early 1960's. It implements various term weighting schemata. It is widely used as an experimentation platform for new term weighting schemata and as the baseline for evaluating other IR models. The Lucene (McCandless et al., 2010) system is an open source information retrieval library. It implements

document indexing and integrates different IR models such as binary and vector space model. Lucene is widely adopted in IR experiments and in TREC competitions, such as in Lin (2006), Cohen et al. (2007) and Alhadi et al. (2011).

Distance (Q, D)	Binary Term Vectors	Weighted Term Vectors
Inner product	$ Q \cap D $	$\sum_{i=1}^n q_i d_i$
Dice coefficient	$\frac{2 Q \cap D }{ Q + D }$	$\frac{2\sum_{i=1}^n q_i d_i}{\sum_{i=1}^n d_i^2 + \sum_{i=1}^n q_i^2}$
Cosine coefficient	$\frac{ Q \cap D }{\sqrt{ Q }\sqrt{ D }}$	$\frac{\sum_{i=1}^n q_i d_i}{\sqrt{\sum_{i=1}^n d_i^2} \sqrt{\sum_{i=1}^n q_i^2}}$
Jaccard coefficient	$\frac{ Q \cap D }{\sqrt{ Q + D - Q \cap D }}$	$\frac{\sum_{i=1}^n q_i d_i}{\sum_{i=1}^n d_i^2 + \sum_{i=1}^n q_i^2 - \sum_{i=1}^n q_i d_i}$

TABLE 6-1. VARIOUS SCHEMATA TO CALCULATE VECTOR DISTANCES

LATENT SEMANTIC INDEXING

The classic vector space model introduced above is based on the presumption that each dimension (each term) is independent, which is not true. The latent semantic indexing (LSI) approach presumes that word co-occurrence is a hint of semantic relatedness; it projects correlated terms to the same dimension (Manning and Schütze, 1999: 556) using SVD (singular value decomposition) algorithm. As a result, the representations of documents are mapped into a space with *latent* semantic dimensions and the weight of each dimension for each document is recalculated.

PRINCIPAL COMPONENT ANALYSIS

Another dimension pruning approach is principal component analysis (PCA). Similar to LSI, this approach extracts those axes in the original space which capture as much the variation of the data as possible. In Kim and Zhang (2001), LSI and PCA are also described as approaches to

solving the synonym and the polysemy problems. The assumption is “each document is assumed to be composed of a few topics and each topic has its own vocabulary. The final document is assumed to be generated by a linear mixture of words chosen from topics. In this scheme, we expect that the exact sense of an ambiguous word can be captured by analysing topics which generated the word” (Kim and Zhang, 2001). The “topics” mentioned here are the extracted axes.

6.2.4 PROBABILISTIC MODEL

While the vector space model scores a document based on its similarity to the query, the probabilistic model, instead, scores a document based on the probability for the document to be relevant to the query. As a result, the vector space model focuses on modelling the semantic of a document, in which term weights are calculated based on how important the term is in representing the semantic of the document. In comparison, the probabilistic model approach focuses on building the topic model that the query defines, in which term weights are calculated based on the representativeness of the term to the topic model. The classic probabilistic model is introduced in 1976 by Robertson and Sparck Jones (1976) which is also known as the binary independence retrieval (BIR) model (Baeza-Yates and Ribeiro-Neto, 2011). Below we will introduce the details of the classic probabilistic model.

The probabilistic model calculates $P\left(\frac{R}{D}\right)$, which represents the probability for a given document D to be relevant. Using the Bayes’ theorem³⁴, the formula is transformed to the formula below. Here $P(R)$ represents the average probability for a document to be relevant and $P(\bar{R})$ represents the average probability for a document to be irrelevant; $P\left(\frac{D}{R}\right)$ represents that the probability for a relevant document to be like document D and $P\left(\frac{D}{\bar{R}}\right)$ represents the probability for an irrelevant document to be like document D.

$$\frac{P\left(\frac{D}{R}\right) * P(R)}{P\left(\frac{D}{R}\right) * P(R) + P\left(\frac{D}{\bar{R}}\right) * P(\bar{R})} \quad (6.11)$$

³⁴ Refer to http://en.wikipedia.org/wiki/Bayes%27_theorem

Robertson and Sparck Jones (1976) simplify the model to only focus on terms that are contained in the query. Applying the term independence assumption, $P(D/R)$ could be simplified as the product of the probabilities of individual terms, i.e., $\prod_{ti \in D \cap Q} P(ti/R) \cdot \prod_{tj \in Q - D} P(\bar{tj}/R)$; similarly, $P(\bar{D}/R)$ could be transformed to be $\prod_{ti \in D \cap Q} P(ti/R) \cdot \prod_{tj \in Q - D} P(\bar{tj}/R)$. Here Q represents a user query; ti and tj represent a term.

To calculate the probabilities, the system must first define a subset of documents in the collection to represent the relevant document set R. To do so, they define initial values for the probabilities to retrieve the preferred answer set, which are then refined after rounds of iterations. The initial values are as defined in formula 6.12 and 6.13, where n refers of number of documents that contain term t and N refers to the total number of documents in the set. In each round of iteration, the values are recalculated based on formula 6.14 and 6.15, where N_R refers to number of documents that are relevant and n_r refers to number of documents in N_R that contain term t. The way to calculate $P(\bar{t}/R)$ could be derived from $1 - P(t/R)$ and similar for $P(\bar{t}/R)$.

$$P(t/R) = 0.5 \quad (6.12)$$

$$P(t/R) = n/N \quad (6.13)$$

$$P(t/R) = \frac{n_r + 0.5}{N_R + 1} \quad (6.14)$$

$$P(t/R) = \frac{n - n_r + 0.5}{N - N_R + 1} \quad (6.15)$$

From the above description, we see that there are a few attributes in common between the vector space model and the probabilistic model.

First, they both presume term independence to simplify the representation for queries and documents. Van Rijsbergen (1979) extends the classical probabilistic model to consider term

dependencies. He uses a tree to model the dependencies among terms: the links between two terms represent the dependency relationships and each term could only depend on one other term. The original formula is revised as below.

$$P(D/R) = \prod_{i=1}^n P(t_i/t_{mi}, R) \quad (6.16)$$

In the above formula t_i iterates across all the terms contained in the query and t_{mi} is the node that t_i depends on. For the root node (i.e., node that have no dependency on other terms), $P(t_i/t_{mi}, R)$ will be $P(t_i/R)$. Note that the above formula applies when a document contains t_i and t_{mi} , when it does not, we should use \bar{t}_i and \bar{t}_{mi} instead. Therefore, for each dependency, there will be eight probabilities to estimate, including $P(t_i/t_{mi}, R)$, $P(\bar{t}_i/\bar{t}_{mi}, R)$, $P(t_i/\bar{t}_{mi}, R)$ and $P(\bar{t}_i/t_{mi}, R)$ for R and four corresponding ones for \bar{R} . To build the dependency tree, van Rijsbergen (1979) first uses the expected mutual information measure (EMIM) to measure the strength of the dependencies among each two terms, which represents the weight of the link between the two terms. He then uses the maximum spanning tree algorithm to calculate a tree structure which links all the nodes and maximizes the sum of the weights of the links.

In addition to the term independence assumption, another shared attribute between the probabilistic model and the vector space model is that they both weight a term taking into consideration of its frequency in non-relevant documents. This corresponds to the $P(t/R)$ factor in the probabilistic model and the idf term weighting schema in the vector space model.

BM25 ALGORITHM (ROBERTSON ET AL., 1993)

The BM25 algorithm is one variety of the probabilistic schemata presented in (Robertson et al., 1993, 1996). It has gained much success in TREC competitions and has been adopted by many other TREC participants.

Robertson et al. (1993, 1996) extend classical probabilistic model in several ways. Here we will only introduce one major extension. The classical probabilistic model models only two

statuses for a term in a document, i.e., present or absent; it does not consider the frequency of occurrences. The extended model instead calculates the probability of a certain number of occurrences of a term in a document. Specifically, instead of calculating $P\left(\frac{t}{R}\right)$ and $P\left(\frac{t}{\bar{R}}\right)$, it calculates $P\left(\frac{tf}{R}\right)$ and $P\left(\frac{tf}{\bar{R}}\right)$, where tf represents term frequency. They use the 2-Poisson model to estimate the probabilities. Thus, the original $P\left(\frac{t}{R}\right)$ is modified to be $P\left(\frac{tf}{E,R}\right)P\left(\frac{E}{R}\right)+P\left(\frac{tf}{\bar{E},R}\right)P\left(\frac{\bar{E}}{R}\right)$, where E represents the elite document set and \bar{E} represents the non-elite document set. The formula to calculate $P\left(\frac{tf}{R}\right)$ could be derived in the same way with R being replaced by \bar{R} in the above formula.

6.2.5 BAYSIAN INFERENCE NETWORK MODEL

An alternative model making usage of the probabilistic theory is the Bayesian inference network model. A Bayesian network (Pearl, 1988) is a directed acyclic graph (DAG), within which each node represents a random variable and a link between nodes represents the probabilistic dependency. Turtle and Croft (1991) first apply Bayesian networks to improve document retrieval performance. They propose that the Bayesian inference network model views information retrieval as “an inference or evidential reasoning process in which we estimate the probability that a user’s information need” “is met given a document as ‘evidence’”. An example is given in figure 6-2.

In figure 6-2, each node represents a random variable which has two values: 0 and 1. In the document network part, d_i corresponds to an actual document in the collection, t_j corresponds to a specific text representation of a document and r_k corresponds to a concept. One d_i usually maps to one t_j and vice versa. The concept nodes are extracted from the text representations. In the same document network, the concept nodes could be different types (e.g., single key words or phrases) derived from different concept extraction methods. In the query network part, the symbol ‘ l ’ refers to an information need which could be met by several queries (q_1 to q_n) and c_m corresponds to a concept extracted from the queries. We see that the above model could be simplified to three layers as below by removing the t_j layer, combining the r_k layer with the c_m layer and simplifying the information need to be just one query, as shown in figure 6-3.

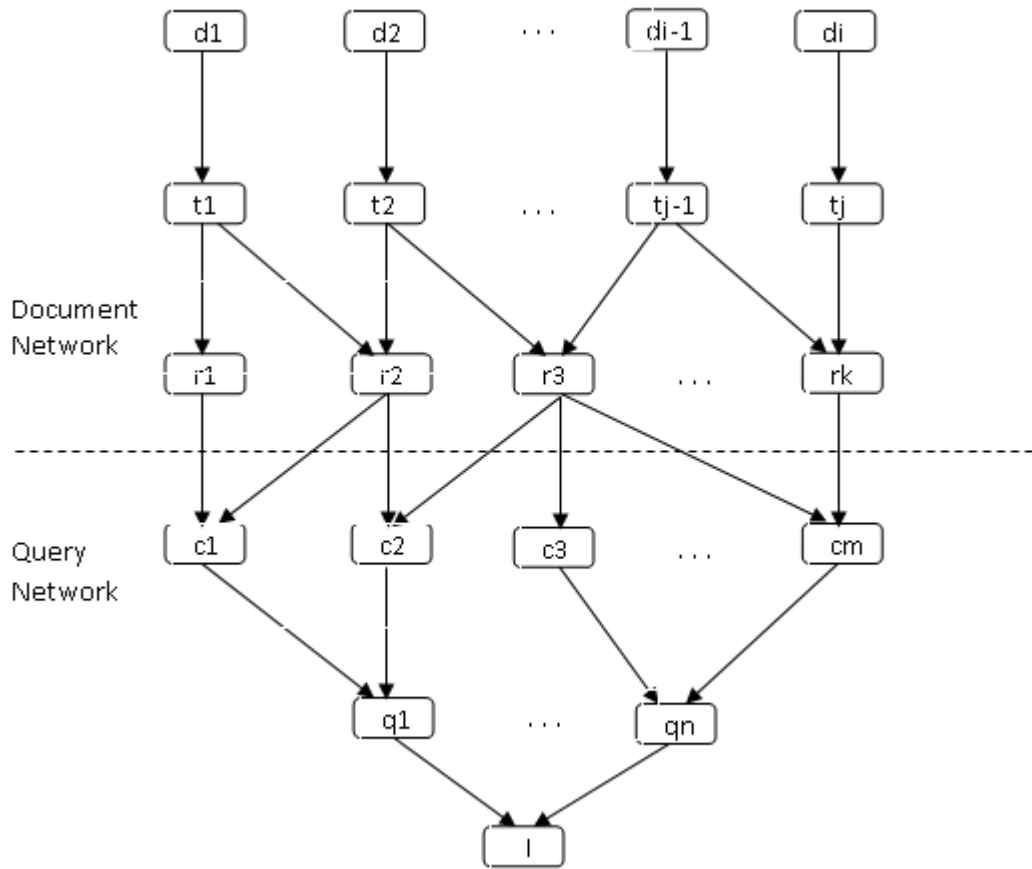


FIGURE 6-2. A BAYESIAN NETWORK MODEL EXAMPLE

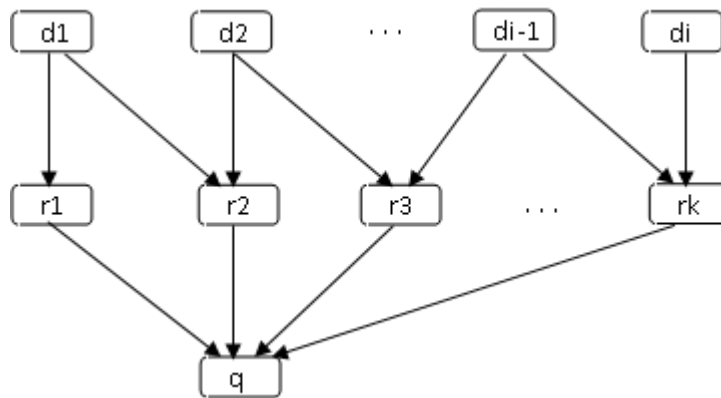


FIGURE 6-3. A SIMPLIFIED BAYESIAN NETWORK MODEL

We have demonstrated in the last section that the probabilistic model scores a document based on $P(R/D)$ – i.e., the probability that a document is relevant. In comparison, the Bayesian network model scores a document d according to “how much evidential support the observation of d provides to query q ” (Baeza-Yates and Ribeiro-Neto, 1999: 51). Specifically,

the ranking is computed as $P(q|d)$, which could be interpreted as the probability that query q could be met given d . The notation is further represented as in the below formula.

$$P(q|d) = \sum_{\forall \vec{r}} P(q|\vec{r}) \times P(\vec{r}|d) \quad (6.17)$$

In the above formula, \vec{r} represents a concept status vector, with the value of each dimension represents the presence of a particular concept r_k ; $P(q|\vec{r})$ means the probability that query q could be met given a concept status vector \vec{r} and $P(\vec{r}|d)$ means the probability of generating \vec{r} given a document d . Suppose there are in total k concepts, then vector \vec{r} would have 2^k values and we would have to calculate $2 * 2^k$ probabilities for one query and document pair. To decrease the complexity of the algorithm, Turtle and Croft (1991) defines four canonical forms of q : not, and, or and weighted sum. The four canonical forms of probability $P(q|d)$ could be directly transformed to the combination of the probability of generating each term, $P(r_1|d)$, $P(r_2|d)$, ..., $P(r_k|d)$, thus greatly simplify the calculation. For example, the canonical form of and – i.e., $P_{and}(q|d)$ – could be simplified as the product of the probability of generating each term, as shown in formula 6.18. The detailed definition of the four canonical forms are as defined in the below formulae.

$$P_{and}(q|d) = \prod_{i=1..k} p(r_i|d) \quad (6.18)$$

$$P_{or}(q|d) = 1 - \prod_{i=1..k} (1 - p(r_i|d)) \quad (6.19)$$

$$P_{sum}(q|d) = \sum_{i=1..k} p(r_i|d) / k \quad (6.20)$$

$$P_{wsum}(q|d) = \sum_{i=1..k} w_i p(r_i|d) / k \quad (6.21)$$

As suggested by Turtle and Croft (1991), the main advantage of the Bayesian network model is that it could be used to combine evidences. As noted in Hiemstra (2001), we could have a Bayesian network which contains two different representations of a document, with each provides different evidence for the document. For instance, one might be the document's title words and the other might be words from the abstract. The model's probabilities might indicate that title words are more important than words from the abstract.

6.2.6 LANGUAGE MODEL

The language model approach is proposed by Ponte and Croft (1998) as a new alternative to traditional vector space models and other probabilistic models. While the probabilistic model inquires about “how probably is it that the document is relevant to the request” – i.e., $P(R/D)$, the language model approach is mainly concerned with “how probable is it that this document generated the request” – i.e., $P(Q/D)$ (Sparck Jones, 2001). Furthermore, it would need to estimate a language model M_d for each document, based on which we could further estimate the probability of generating the request, i.e., $P(Q/M_d)$. If assume term independence, then it would translate to $\prod_{i=1}^m p(q_i/M_d)$. According to Ponte and Croft (1998), the good thing about the language model, compared to the probabilistic model, is that there is no word distributional model assumption over the collection and there is no presumption that a document is a member of a pre-defined class. In contrast, in the 2-Poisson probabilistic model, there is a pre-existing classification of documents into elite and non-elite sets and also a presumption of Poisson distribution of term frequency within each set.

A simple approach to $p(q_i/M_d)$ is tf/doc_length , that is, the frequency of occurrences of the term q_i in the document normalised by document length. The major problem with this formula is that the probability will be zero for a document where the query term is missing. Therefore, smoothing is one major issue for modelling $p(q_i/M_d)$. To do so, Ponte and Croft (1998) use cft/cs as the default value when a document does not contain the term, where cft is the frequency of occurrences of the term in the whole document collection and cs is the size of the document collection. Furthermore, only use one data sample (i.e., q_i in document d) to estimate $p(q_i/M_d)$ is not robust enough. So they plug into the original formula a factor of $p_{avg}(q_i)$, which calculates the average $p(q_i/M_d)$ in all documents that contain term q_i . The final formula is a weighted sum of the original $p(q_i/M_d)$ and $p_{avg}(q_i)$. The weight is calculated based on term frequency of q_i in d ; the underlying thinking is that when q_i gets

really big then the influence of the general model should be smaller. They experiment on the TREC datasets and obtains significant improvement over a classical tf-idf approach.

Different variations for the initial language model are proposed later on. One major direction of improvement is to plug in a translation model between the query and the document, such as in Berger and Lafferty (1999) and in Jin et al. (2002). Berger and Lafferty (1999) presume that when a user forms a query from an information need, s/he will imagine the ideal documents and then translates these documents into a query. Therefore, for a retrieval system to retrieve the best document, it must model this translation process. Thus,

$$p\left(\frac{q_i}{d}\right) = \sum_w t(q_i | w) l(w | d),$$

where $t(q_i | w)$ is the translation model and $l(w | d)$ is the

document language model. As it is hard to obtain a training corpus large enough to estimate $t(q_i | w)$, they further generate some synthesized queries for each document to represent the document content, thus decrease the choice of w . The synthesized queries are generated by choosing words that would have the maximum mutual information statistics $I(w, d)$, which is calculated by $p(w, d) \log\left(\frac{p(w | d)}{p(w | D)}\right)$. They used the expectation maximization (EM) algorithm to estimate t based on the training corpus. Their experiments on the TREC datasets demonstrate substantial improvement over standard vector space model. Similarly, Jin et al. (2002) presume that the titles of documents work similar to queries and therefore estimate the translation model using title and document pairs as the training corpus. They experiment on TREC AP88 (Associated Press, 1988), WSJ90-92 (wall street journal from 1990 to 1992) and SJM (San Jose Mercury News, 1991) and the results show that their method outperforms both the classical language model and the Okapi system (probabilistic model).

It is worth noting here that the translation model also provides a smooth function since it allows the document to not contain all (or even any) the query words to derive a non-zero score.

6.2.7 LIMITATION OF IR MODELS FOR RETRIEVING EXTENDED TOPIC

The above-described key-word-matching-based IR technology has many limitations. In particular, when applied to retrieve extended topics, it might be less effective in retrieving the generic part than the specific part.

The above-mentioned IR models score a document according to how many key words matches could be found in the document. Some hidden or related concepts would not be identified. For example, for a query asking about the colour of something, we would not know

that red is a type of colour and therefore a document containing word ‘red’ is possibly relevant. In previous chapters (section 3.2, 4.3 and 5.6), we noted that the generic part of an extended topic (i.e., the generic topic) confines the required information and will be replaced in the answer by details that are *unknown* to the questioner; in comparison, the specific part sets up the context of interests and will be kept as the *given* information in the answer. Therefore, to retrieve information relevant to a generic topic, the system has to model the relation it has with the unknown details in the answer. Typical IR techniques clearly do not meet this requirement. For example, passage [6.1] talks about the climate of Fiji; the topic climate cannot be easily detected since word ‘climate’ does not appear in the text. One simple solution to this problem is to build a list of related key words to expand the query. This technique is called query expansion and is commonly adopted in current IR systems. Related key words are constructed using a thesaurus or using simple statistical-based method. In section 4.4, we have adopted a similar approach to detect the connection between the generic topic and the new information in a discourse. Specifically, we built the signatures of a generic topic (a list of related key words) based on word co-occurrences. The signatures of the topic climate include ‘temperature’, ‘annual’ and ‘average’, etc³⁵. Since passage [6.1] contains many signatures of climate, as highlighted in grey, query expansion technique could retrieve this passage for the topic climate. We will discuss detailed query expansion techniques and their limitations in section 6.3.

[6.1] At Suva the average summer high temperature is 85 F (29 C) and the average winter low is 68 F (20 C); temperatures typically are lower in elevated inland areas. All districts receive the greatest amount of rainfall in the season from November through March, during which time hurricanes are also experienced perhaps once every two years. While rainfall is reduced in the east of the larger islands from April to October, giving an annual average of 120 inches (3,050 millimetres) per year, it virtually ceases in the west, to give an annual rainfall of 70 inches, thus making for a sharp contrast in both climatic conditions and agriculture between east and west.³⁶

As noted in section 4.2, many generic concepts specify one relationship or work as shorthand for a family of relationships. For instance, the concept ‘procedure’ implies relationships like

³⁵ The full list is included in appendix D.2.

³⁶ Cited from <http://www.britannica.com/EBchecked/topic/206686/Fiji/53915/Climate>

goal, precondition, instrument and method. Such relationships are usually conveyed in cue phrases, discourse connectives and prepositions. For example, the phrase ‘in order to’ conveys a purpose relationship. However, this sense is no longer there when it is broken into ‘in’, ‘order’ and ‘to’ in the typical key word based retrieval approach. For example, sentence [6.2] contains all of the three words but it does not contain a purpose relationship; so as in passage [6.3], which contains many matches to word ‘in’, ‘order’ and ‘to’ respectively, but none of them conveys a purpose relationship. In addition, many cue phrases or prepositions could express different relationships under different circumstances. For example, ‘with’ could indicate the companion relationship as well as the instruments used to achieve a goal or perform an action; ‘to’ could indicate purpose relationship or sequential relationship between two actions. Therefore, when separating these phrases or prepositions out of the context, as in typical IR models which treat each indexing unit independently, they become too ambiguous to retrieve good results.

[6.2] “Libya council seeks to restore order in Tripoli”³⁷,

[6.3] “President George W Bush last night ordered American troops to be moved to the coast of war-ravaged Liberia in preparation for a landing in support of West African peacekeeping forces.

The move had been bitterly opposed by Pentagon officials who said it could lead to the type of humiliation suffered in Somalia in 1993 which has haunted military planners since.

The White House did not specify the number of US soldiers involved, but the Pentagon said the assault ship Iwo Jima, leading a three-ship group carrying 2,300 marines, had entered the Mediterranean en route to Liberia.

"We're working very closely with the United Nations," Mr Bush said. "They will be responsible for finding a political solution and they will be responsible for relieving US troops in short order."

The US decision came as Liberian rebels battling forces loyal to President Charles Taylor announced another ceasefire, their third in recent weeks of fighting that has killed hundreds and produced some of

³⁷ Cited from <http://www.bbc.co.uk/news/world-africa-14770357>

the worst scenes in a decade of civil war. The peacekeeping force, led by Nigeria, is still arguing about how and when to enter the country.”³⁸

Above says that the key-word-matching-based IR technology may not be effective in retrieving the generic part of an extended topic. Furthermore, here we will show that *directly* applying the generic part to match retrieve documents may even bring noise when using some of the IR ranking models.

In the vector space model approach, tf-idf is the most applied term weight schema. The underlying thinking of using idf is that some terms may occur in a broader context and therefore one match to such a term may not be a strong evidence for the document to be relevant. In other words, such terms are relatively weak in its differentiating ability. Idf models the broadness based on terms' document distribution. This schema does not work well for the generic part of an extended topic. This is due to the fact that generic concepts (i.e., concepts that are typically used in the generic parts of extended topics) are often implicit in documents. As a result, the total number of times a generic concept occurs in a document collection is smaller than the times that it actually exists; in other words, the idf value of the generic concept is larger than the actual value. Therefore, when using the vector space model for document ranking, the generic parts will be over-emphasized and will bring noise. This conclusion also applies to the extended Boolean model for the same reason.

The inference network model scores documents based on $P(q|d) = \sum_{\forall \vec{r}} P(q|\vec{r}) \times P(\vec{r}|d)$, where \vec{r} is a concept vector. $P(r_i|d)$ is further defined in Turtle and Croft (1991) as $\alpha + \beta * tf_i + \gamma * idf_i + \delta * tf_i * idf_i$. We see that here idf is again applied as an important parameter for scoring documents. As a result, the above-mentioned problem also exists when using the inference network model for ranking retrieval.

The language model approach ranks a document based on $\prod_{i=1}^m p\left(\frac{q_i}{Md}\right)$. It uses $\frac{cft}{cs}$ as the default value when a document does not contain term t, where cft is the frequency of occurrences of the term in the whole document collection and cs is the size of the document collection. The cft of a generic concept is smaller than the actual value. Therefore, a document

³⁸ Cited from <http://www.telegraph.co.uk/news/worldnews/africaandindianocean/liberia/1437217/US-orders-marines-to-Liberia.html>

that does not contain the generic concept will be scored smaller than actual. Again it to some extent over-emphasizes the importance of containing generic concepts.

Only the probabilistic model seems not to have a bias towards generic concepts. The key reason is that it uses an ideal document set as a reference rather than trying to weight the importance of a concept based on term distributions. Specifically, it scores a document based on $P(t/R)$ and $P(t/\bar{R})$. To judge whether a document belongs to the relevant set, it will use the term distribution in the ir-/relevant document sets as a base, the basic assumption being that the ir-/relevant documents resemble each other in terms of term distributions.

Above we point out three problems when applying typical IR approach to retrieve extended topic, including generic concepts might be hidden in a relevant document, the key word based indexing method breaks a phrase into meaningless units and directly using the generic part to match retrieve documents may bring noise. These problems indicate that deeper natural language processing (NLP) techniques should be involved to construct a more sophisticated representation of documents in comparison to just a list of key words. In the history of the IR research, various NLP techniques have applied to enhance IR result, such as phrasal retrieval, information extraction and semantic analysis. However, except for a few shallow linguistic analysis techniques, such as using a stemmer to remove word suffixes, which are broadly applied in IR systems, in most of the areas we could not draw a confident conclusion that NLP technology could be applied to in a large scale to improve the performance of an IR system. Statistical approach gains popularity in the early 90's and remains as the mainstream IR approach till today. As Sparck Jones (2003) points out, "though continued attempts were made to show that 'proper' language processing, i.e. syntactic and possibly also semantic parsing, was required for better retrieval performance, this was not supported by the test results". A few key reasons are: linguistic analysis is not accurate enough and the errors generated are further introduced into the IR system; linguistic analysis is often too expensive to apply to a large scale and is less robust compared to statistical approaches based on simple key words; word frequencies and co-occurrences already implicitly encode the information that linguistic analysis aims to extract; current IR experiment setup does not stress the problems that linguistic analysis aims to solve; etc. We will introduce some of these technologies later in this chapter.

We also pointed out the query expansion techniques could alleviate the first problem and are commonly used in the typical key-word-matching-based IR systems; below we will briefly introduce the query expansion technique.

6.3 QUERY EXPANSION

6.3.1 QUERY EXPANSION TECHNIQUES

Often users have difficulty in formulating good queries. First, most users do not know how a query is used in an IR system. Secondly, the match retrieval method requires the query to be similar to a certain degree to the relevant documents. This seems to be contradictory as it requires users to formulate something that are similar to what they are looking for. In reality user queries are usually very short. Specifically, the average length of queries to a search engine is just over two words (Sparck Jones et al., 2007). To mitigate this problem, query expansion technique is typically used in IR systems to extend the original query with a list of related words so as to retrieve more relevant documents.

To identify terms that are related to the original query, one approach is to directly use a dictionary or a thesaurus. For example, Neumann et al. (2003) present a cross-language IR system which use synonyms defined in EuroWordNet to expand the original query.

There are also approaches to automatically learning the correlation among words based on word distribution. Two basic types of strategies could be attempted: global ones and local ones (Baeza-Yates and Ribeiro-Neto, 1999).

“In a global strategy, all documents in the collection are used to determine a global thesaurus-like structure which defines term relationships.” (Baeza-Yates and Ribeiro-Neto, 1999) There are different algorithms used to calculate the thesaurus. In one of the algorithms introduced in (Baeza-Yates and Ribeiro-Neto, 1999), the strength of relationship between two words is calculated based on term co-occurrence over the document collection.

“In a local strategy, the documents retrieved for a given query q are examined at query time to determine terms for query expansion.” (Baeza-Yates and Ribeiro-Neto, 1999: 123) One simple approach is just to add the most frequent words in the top ranked documents or relevant documents. The other approach is to design algorithms to calculate word co-occurrence in the retrieved relevant documents and then add the strongly correlated ones.

Both of two approaches do not consider the comparison between relevant documents and irrelevant documents. Another approach is to find out the terms that are largely shared among relevant documents and less shared among the irrelevant ones. One of the formulae is given by Rocchio (1971), as shown below.

$$\text{standard_Rocchio: } \vec{q}_m = \alpha \vec{q} + \frac{\beta}{|D_r|} \sum_{\forall d_j \in D_r} \vec{d}_j - \frac{\gamma}{|D_n|} \sum_{\forall d_j \in D_n} \vec{d}_j \quad (6.22)$$

Where D_r is the collection of relevant documents and D_n is the collection of irrelevant documents. $|D_r|$ and $|D_n|$ are the number of documents in each collection.

While most of the above-mentioned approaches derive an explicit set of related query terms to expand the original query, there are also approaches which have similar effect but are initially designed to tackle different problems. For example, one of the smoothing strategies in the language model approach is to introduce a translation model. Instead of directly using the query words to match retrieve documents, the translation model calculates the probability for translating a word in a document to one of the word in the query, i.e., $t(q_i/w)$. Details are described in section 6.2.6. We see that, similar to query expansion, the translation model could potentially find relevant documents that contain none of the query words.

6.3.2 LIMITATION OF QUERY EXPANSION

Query expansion may partly solve the first problem mentioned in section 6.2.7, which is, some concepts may be hidden in a relevant document. However, query expansion is still a keyword-matching-based approach, thus the second problem (i.e., the key-word-based-indexing method breaks a phrase into meaningless units and separates an ambiguous phrase/preposition from its context) still remains. Besides, query expansion still use the generic part of extended topic to retrieve documents (at least in the first round); thus it also does not solve the third problem mentioned above (i.e., directly using the generic part to match retrieve documents may bring noise).

Another problem with query expansion is that it may only retrieve documents similar to the top results retrieved in the first round and thus reduce the diversity of the results. The reason is that the original query is expanded by terms extracted from the top results retrieved in the first round. For example, if we look for sources of pollution of Beijing and if many of the top results talk about factories of heavy industry, then key words ‘heavy’, ‘industry’ and ‘factory’

would be added to expand the original queries; as a result, other documents talking about ‘daily transportation’ and ‘construction work’ may not be retrieved.

6.4 APPLY EXTENDED TOPIC TO IR

In section 6.2.2 we point out three problems in applying typical IR techniques to retrieve extended topics. In section 6.3 we introduce query expansion techniques which could partly solve the first problem but still have a lot of limitations. There are also IR systems that integrate advanced NLP techniques; however, they do not differentiate between the generic part and the specific part of extended topic. Below we will present our solutions for the problems.

In section 4.2 we analyse WH-questions and classify extended topics into four classes. Here we propose different solutions for different classes of extended topic.

In the first class of extended topic, the generic part denotes a simple attribute, e.g., colour and name. The name of the attribute in a question is replaced by attribute values in answers. If the values of the attribute are enumerable, we could model the match between the attribute name and the attribute values using a thesaurus or an ontology. For example, we could easily encode the relationship between the attribute name ‘colour’ and concrete colours such as ‘red’, ‘green’ and ‘blue’ in a thesaurus or an ontology. Thus, the above problem could be solved by using the query expansion technique to append a list of concrete attribute values extracted from a thesaurus or an ontology. Some attributes such as dates and numbers are not enumerable, simple pattern matching or shallow NLP technology could be used to pre-process the documents to identify these attributes. For example, Srihari and Li (2000) combines manually defined patterns and automatic machine learning method (e.g., maximum entropy and hidden Markov chain) to identify named entities, including person names, organisation names, dates and times, etc. The study of extracting named entities falls into the research area of information extraction (IE), which is applied in a question answering (QA) system to retrieve and extract answers for questions asking about simple attributes such as dates and numbers. We will introduce QA in section 6.7. There is also research on building a formal knowledge representation schema and conducting deep natural language processing (NLP) to extract knowledge to fill this schema. For example, Noy (1997) defines an ontology in medical domain which uses a frame formalism to represent concepts, attributes and medical processes. NLP technologies are used to process texts to identify the associated frames. In this process, attribute values will be extracted to fill the predefined attribute slots. In the

document retrieval stage, queries could be matched against the attribute slots and the attribute values will be retrieved. As NLP technologies are too expensive to apply to a large document collection, mainstream IR research still centres on shallow statistical based approaches. Since there are already many techniques targeting at the first class of extended topic, the focus of the thesis will be on the remaining three classes of extended topic.

The second and the third classes of questions retrieve an entity/proposition/event based on its relationship to the focused entity. We will therefore focus on detecting the relationship. As mentioned in section 5.6, to detect a particular relationship, we could collect a list of phrases/prepositions/discourse connectives often being used to express the relationship. In the document retrieval stage, we will replace the name of the relationship in the original query by the phrase list to retrieve relevant documents. For example, for a question ‘what is the purpose of aspirin’, we could use ‘be used for’, ‘helps to’ and many other similar phrases to replace ‘purpose’. While most IR systems use keywords to match retrieve documents, phrasal retrieval (i.e., using phrases to match retrieve documents) has also been experimented with. We will introduce the literature on phrasal retrieval in section 6.5. Further, as the second problem presented in section 6.2.2 points out, such phrases/prepositions/discourse connectives may become very ambiguous when separated out of the context. Therefore, directly using them to retrieve documents may bring a lot of noise. To resolve this problem, we suggest considering the association between the generic part and the specific part. Specifically, when scoring a document, the system only counts a relationship phrase (occurred in the document) as a match (to the generic part) if it is highly likely to be associated with the inquired specific entity. In other words, a phrase must occur together with the focused entity in a document to be count as a match. For example, for question ‘what is the purpose of aspirin’, the cue phrase ‘be used for’ must occur together with the focused entity ‘aspirin’ to be count as a match. This thinking of ensuring the association between different elements of a query is also similar to phrasal retrieval.

The above proposed method for the second and the third classes of questions will be applied in chapter 7 to retrieve relevant documents for causal queries, i.e. questions asking about the cause of some specific event/act/state/phenomenon. Causal questions are the most frequent question type based on the study on WH-questions presented in chapter 4. The answer to a causal question, i.e., the cause, is not required to fall into a certain semantic class. Indeed, almost anything could be a cause of some other things. Thus, instead of detecting the cause itself, the key lies in detecting the cues signifying the causal relationship. Details about how to

detect the causal relationship and how to ensure it is related to the specific event/act/state/phenomenon in question will be discussed in chapter 7.

For an extended topic of the fourth class, the concept of the generic part primes a category of relationships/facts. For example, a description of the anatomy of an organ would typically be a short passage which contains a lot of compositional and positional relationships. Again, there are typical phrases to express such relationships, for instance, 'consist of' and 'locate at'. However, different from the above-mentioned second and third classes of topics, here we cannot directly apply phrasal retrieval. This is because it is not sufficient to tell whether a passage is about anatomy by a single phrase like 'consist of'. Instead, we will test whether the passage contains a sufficient amount of compositional and positional relationships. We suggest using text categorisation approach to automatically classify a document to be relevant or irrelevant. Instead of outputting a simple category, most classification models output a degree of confidence for a document to fall into a certain category, which could provide the base for us to score document relevancy. Text categorisation (TC) has been widely adopted in current IR systems, which will be introduced in section 6.6. The novelty of our approach lies in separating a question into two parts and identifying the necessity of applying TC techniques to one part and using simple word matching techniques to another part. Here comes a problem of how to combine the score of the two scoring results. We suggest dividing the retrieval process into two stages:

- a. In the first stage, use typical IR approaches for retrieving documents that are relevant to the specific part;
- b. In the second stage, use a text categorisation approach to re-rank the retrieved documents according to degree of relevancy to the generic concept.

We could certainly think of other ways to combine the two scores. However, the two-stage approach presented above has several benefits:

- a. The classification approach is much more expensive than a simple key word matching approach; therefore, only apply the expensive method to a few top ranked documents (with high likelihood of being relevant) saves a lot of time compared to applying it to the whole corpus.
- b. In general, a document that ranked high according to the generic part is less likely to be relevant than a document that ranked high according to the specific part. In other words, again, using the generic part as the prime criteria will bring a lot of noise.

Therefore, we only apply the generic concept to retrieve documents at the second stage after we are pretty confident that the document is relevant to the specific part.

The above two-stage process resembles a QA system, which also adopts a two-stage approach, a brief introduction of QA system will be presented in section 6.8.

In chapter 8, we will apply the two-stage approach to retrieving relevant documents for procedural and biographical questions. Procedural questions retrieve a list of steps for achieving a specific goal. Biographical questions retrieve biographical facts such as profession, birthdates and education, etc. We will introduce the details about how we train document classifiers to recognise document talking about procedures and biographies and how we combine the classification results into an IR system.

It is worth noting that all of the above-proposed methods presume that the generic part and the specific part of extended topic are already marked out; automatically detecting the generic part and the specific part is out of the scope of this thesis.

6.5 PHRASAL RETRIEVAL FOR IR

Although most IR systems use key words as the basic units to represent queries and documents, however, there are also substantial amount of studies on phrasal retrieval, i.e., identify and extract phrases to represent queries and documents. The benefit of phrasal retrieval could be summarised in a few points. First, a single word by itself may denote a concept totally different from it would refer to when being combined with several other words. For example, 'space needle' refers to a tower in Seattle and cannot be broken into 'space' and 'needle'. Secondly, a single word may have several different senses and using phrase retrieval, by adding some contexts (basically other key words), helps in disambiguation. For example, 'bank' will be disambiguated when we use 'British Bank' to match retrieve documents. Last but not least, using phrases ensures the relationship between different terms is kept in the same way in the document as in the query. For example, the relationship between 'history' and 'Britain' in the phrase 'history of Britain' may not be maintained in a document that contain both 'history' and 'Britain' in a non-consecutive manner.

Most IR indexing formulae “weights the importance of a term to a document in terms of occurrence considerations only, thereby deeming of null importance the order in which the terms themselves occur in the document

and the syntactic role they play; in other words, the semantics of a document is reduced to the lexical semantics of the terms that occur in it, thereby disregarding the issue of compositional semantics.” (Sebastiani, 1999)

“When a user presents an information need in the form of a natural language query, they specify, by their choice of particular linguistic relationships, a variety of meaning connections between the words in the query, Treating such query as a set of weighted terms ignores such connections. One of the advantages of a structured query may be to capture, some of the relational structure normally expressed in natural language.” (Croft et al., 1991)

There are many studies comparing between phrasal retrieval and key word based retrieval. The conclusions seem not to be in agreement with each other: with some studies (Mitra et al., 1997; Carpenter, 2004) show that phrase retrieval does not outperform key word based retrieval; but some others (Mishne and de Rijke, 2005; Croft et al., 1991) conclude the opposite. A general impression derived from the above work is that phrasal retrieval has to be flexible to allow different level of evidences to be considered, for example, the status of only containing part of the component words, the status of containing all component words within a certain distance, etc. Recent studies such as Svore et al. (2010) and Song et al. (2009) design method to calculate the *importance* of phrases in helping information retrieval, which is plugged in into the document ranking algorithm. Their experiments show great improvement over key-word-based and basic phrasal based retrieval approaches. We will introduce their work later in this section.

“If a sufficient number of phrase descriptors are assigned to documents and queries, and if these descriptors are predominantly good indicators of document content and information need, then substantial improvements can be achieved.”(Fagan, 1987)

“On average, the use of phrases does not significantly affect precision at the top ranks. Preliminary observations indicate that phrases are more useful in determining the relative ranks of low-ranked documents. Phrases are useful for some queries, but not others. The major issue seems to be one of query accuracy versus query coverage”. (Mitra et al., 1997)

There are two major approaches in identifying phrases: syntactic phrases (identified by syntactic analysis) and statistical phrases (identified by statistical analysis method). Statistical

approach for phrase identification is to examine word co-occurrence patterns, e.g., co-occurrence of terms in documents at a rate greater than that would be expected by random chance (Greengrass, 2001). Co-occurrence could be combined with adjacency, e.g., if two (or more) terms co-occur within a few words of each other at a rate greater than chance, the probability that they are related semantically increases further (Greengrass, 2001). When comparing between the performance of syntactic phrases and statistical phrases, most studies indicate that syntactic phrases have high cost but do not bring any gain over statistical phrases.

“With regard to the relative value of non-syntactic and syntactic phrase indexing, the precision averages show that non-syntactic phrase indexing is significantly better than syntactic phrase indexing for the CACM collection,”... “Examination of individual queries, however, shows that there is great variability in the performance of both syntactic and non-syntactic phrase indexing.” (Fagan, 1987)

“Insofar as compound index terms, as opposed to single words, were of use, so-called statistical phrases defined by repeated word tuples were just as effective as ones obtained by explicit parsing (Salton, et al., 1990, Croft et al. 1991).” (cited from (Sparck Jones, 2003))

“Lewis, 1992 and 1996 suggest that statistical weighting techniques should be applied to phrase descriptors, even if they are generated by NLP or combined NLP/statistical techniques.” (cited from (Greengrass, 2001))

“When phrase matches alone are used to rank documents, syntactic phrases perform better than statistical phrases, but this advantage disappears when single terms are used in indexing and retrieval;...” (Mitra et al., 1997)

One major problem of syntactic phrases, as indicated in (Fagan, 1987), is its restrictiveness in phrase selecting. In comparison, statistical approach is freer and more robust. However, the weakness of the syntactic phrases is also an advantage; compared to statistical phrases, they will also avoid a lot of noise brought by statistical analysis. Besides, syntactic method could identify relationships between words at fairly long distances, which is also an advantage that statistical approaches lack. Fagan (1987) further propose to combine the two methods together.

Below we will introduce, in details, a few studies on phrasal retrieval.

Croft et al. (1991) design experiments to verify a) “whether structured queries incorporating phrases will be more effective than unstructured queries” and b) “phrases selected automatically will perform as well as phrases selected manually”.

They first point out that “despite the significant amount of work on phrases”, “the relationship of phrases to the retrieval model has not been sufficiently examined”. In more concrete terms, it is not clear whether a phrase should “be treated as an index terms, similar to index terms derived from single words”, or should “be treated as a relationship between index terms”. They propose four inference net models based on different views of phrases. For example, the first model views a phrase independent of the component words, thus, the belief of the presence of a phrase is not calculated based on the belief of the presence of each component words; in the second model, a phrase is viewed as the concatenation of the component words, thus, the belief of the presence of the phrase concept is a combination of the belief of the component words; etc.

Three concrete retrieval approaches are derived from the first and the second model. One is a conjunctive approach based on the second model. In this approach, the belief of the presence of a phrase is the production of the beliefs of the component terms and the score of a document is the mean of the belief of a phrase and other remaining terms in the query. Another one is a proximity phrase approach based on the first model. In this approach, the presence of phrase could only be sufficiently evidenced by the occurrence of the component terms within a certain distance. In other words, component words must occur within a certain distance to count as a match. The hybrid approach combines the first two, in which the belief of the presence of a phrase depends on individual terms as well as the frequency of a phrase.

They try with several different approaches to automatically extract phrases, including using a partial parser, using a stochastic tagger and using phrases from a dictionary or thesaurus or combinations of the above approaches.

Experiments results show that phrasal retrieval (including both manually produced phrases and automatically generated phrases) do improve the single term based retrieval. Moreover, there is little difference between manually selected phrases and automatically generated phrase using syntactic analysis combined with various pruning and filtering method.

Lewis and Sparck Jones (1996) note that with the availability of large document database, natural language indexing will take over the role of conventional control language indexing. While simple natural language indexing techniques, such as key word based indexing, have

been shown adequate in a wide range of experiments, however, they have not been tested in a very large scale. In their work, they discuss how NLP-based approaches could help to improve the indexing method and whether at the current stage automatic NLP techniques reach the required maturity to process full texts in a large scale to be able to help improve retrieval result. They provide specific suggestions with regard to the indexing unit and the matching methods in the searching process.

In terms of the indexing units, they suggest using linguistically solid compounds or basic propositions. They emphasize that statistical weighting should be applied to these linguistic constructs. Finally, “Compound units of this time should not be further combined into frames, templates, or other structured units” due to the high-cost to build complex structures.

When it comes to the matching method, they suggest allowing uncertain evidences to be considered. “compound terms will not be identified as definitely occurring or not occurring in a document. Rather, each document will provide some amount of evidence for the presence of each known concept”. This means that to match the concept “prefabricated units”, the verb phrase “(they) prefabricated units” will provide a certain degree of evidence to this concept. And the occurrence of “prefabricated” and “units” separately will still provide some weak evidence. They further propose to conduct phrase normalisation to allow non-identical phrases to match to each other, for example, using a word stemmer to remove word suffixes or using a thesaurus to relate synonyms.

Mitra et al. (1997) investigate whether additional improvements could be obtained by using phrases in indexing and retrieval given a good basic document ranking scheme. They also experiment to verify whether there is significant difference in the benefits obtained from syntactic phrases and from statistical phrases. They derive statistical phrases by extracting all pairs of non-function words that occur contiguously in at least 25 documents in disk 1 of the TREC collection; to extract syntactic phrases, they use the Brill POS tagger (Brill, 1993; Brill, 1994) to preprocess the documents and identify certain POS tag patterns as noun phrases. Since applying POS tagging to all documents is rather expensive, so instead of processing all documents to identify the phrases, they first use a term weighting schema to pre-rank the document and then extract syntactic phrases within the top 100 documents only.

They use a variation of the standard vector space model for ranking. To derive the idf for a phrase, they use different schema to derive the phrase idf based on the idf of each word constituent, such as, the maximum of the idf of each component word, the minimum of each component word, the arithmetic mean of each component word, etc.

They experiment on TREC queries and the TREC document collections consisting of the Wall street Journal, AP Newswire, and Ziff-Davis sections of TREC disk 2.

They compare among different approaches: a) terms ranking only; b) phrase combines with term on the top 100 documents and c) phrase ranking only on the top 100 documents. The experiments result show that: b) does not clearly outperform a); c) performs much worse than a). When comparing between syntactic phrases and statistical phrases, under the settings of b), there is no clear difference between syntactic phrase and statistical phrase, under the settings of c), syntactic phrase is slightly better.

They conclude that: a) “when phrase matches alone are used to rank documents, syntactic phrases perform better than statistical phrases, but this advantage disappears when single terms are used in indexing and retrieval”; b) “on average, the use of phrases does not significantly affect precision at the top ranks”; c) Observations show that phrases perform well in some of the queries, but not others. Phrases seem to be more useful when single terms retrieval does not perform well.

Mishne and de Rijke (2005) experiment on using phrases to improve Web documents retrieval result. They follow the experiment setup of the web tracks at TREC 2003 and TREC 2004 and experiment on 1.25M documents from .gov domains.

In the experiments, they consider all the word n-grams in a query as phrases and do not use statistical or syntactic analysis to refine the set. They apply several different methods to match a query phrase: one exact phrase matching method which only counts strict word n-grams in a document as a match; one proximity method which counts as a match if all the words in the query phrase are within a certain distance (window) in a document. Here we call the match in the proximity method as a proximity phrase. There are again two variations in the proximity method, including a fixed window approach and a varied window approach. In the fixed window approach, any proximity phrase within the pre-set distance (window) would be counted as one match and would be weighted the same with other proximity phrases. In the varied window approach, a proximity phrase with distance x will be count as $x-x+1$ matches; in this way, the proximity phrase with short distance will be rewarded.. The experiment result shows that a) all phrase methods outperform the baseline, i.e., single term matching approach; b) fixed window approaches outperform the exact phrase matching approaches, but are generally not as good as the flexible proximity terms.

Song et al. (2009) model the *compositional power* and the *discriminative power* of a phrase to improve the performance of a document retrieval system. The *compositional power* of a phrase refers to how likely a phrase can be represented as its constituent words without forming a phrase. They apply a bunch of features to model the compositional power such as the average no. of occurrences of a phrase in the document collection and whether a phrase only occurs once in a document, etc. The *discriminative power* of a phrase is to measure whether the constituent words or co-occurrences of the constituent words are discriminative enough for retrieving relevant documents. Again, there are a bunch of features applied to measure the discriminative power of the constituent words, including their document frequencies and the ratio of the times they form phrase to the times that they co-occur but do not form a phrase. A RankNet model (Burges et al., 2005) is used to combine these different features to compute a general phrase importance score.

Their method shows great improvement over a standard word unigram-based language model and phrase-based language model approach.

Svore et al. (2010) explores whether using flexible proximity information (i.e., spans) can improve web retrieval accuracy. A span must start and end with a query term and must match a certain length limitation. Different from n-grams, query terms contained in a span do not need to keep their original order in the query. They define a method to score the “goodness” of a span in retrieving relevant documents based on the presence of third-party phrasal information within spans, the formatting and structure of spans, the density of query terms in the span and so on. The third-party phrasal information are extracted from Wikipedia titles and by mining search engine’s query logs for common n-gram occurrences. They use LambdaRank (Burges et al., 2006), a neural-network-based machine learning model, to combine these different features to compute the goodness score.

They compare among different ranking approaches: using simple term unigrams, using adjacent bigrams, using both adjacent and non-adjacent bigrams, using simple span features (e.g., span length and frequencies) and using span goodness score. The performance of the different approaches are in the same order as listed above, with term unigrams performs the worst and the span goodness score performs the best.

6.6 TC AND TC FOR IR

Text categorisation (TC) has close kinship to information retrieval. Sabastiani (2005) defines TC to be the task of classifying documents from a domain D into a given, fixed set $c=\{c_1, \dots, c_m\}$ of

pre-defined categories (aka., classes or labels). In theory, IR could be seen as the task of classifying documents into a relevant set and an irrelevant set, therefore, the problem of IR is essentially the same to the problem of TC. However, in practice, TC is usually applied to slightly different tasks from IR. Specifically, an IR model is a general method for calculating the degree of relatedness between a document and an information need; therefore, it is usually applied to many different queries against a relatively stable document collection. Instead, in a TC system, a complex model is generated to characterise one or a few steady information needs. There are constantly incoming documents being processed by this model to identify relevant documents.

TC has a long history, dating back to at least the early 60's. However, until the late 80's, the most effective approach to TC seemed to be that of manually building automatic TC systems by means of knowledge-engineering techniques (Sebastiani, 1999). In the 90's this perspective has been overturned, and the machine learning paradigm to automated TC has emerged and definitely superseded the knowledge-engineering approach (Sebastiani, 1999).

The basic components of a typical current text classification problem are the corpus, the categories, the feature being used and the machine learning models. There are several corpora widely used in most current TC research, including, Reuters-21578, Reuters-22171, Reuters Corpus Volume I and II, the TREC-AP corpus, TIPSTER corpus, 20 Newsgroups and OHSUMED. Stricker et al. (2000) experiment on several news resources to find news addressing specific topics. They present a method for automatically generating "discriminant terms" (Stricker et al., 2000) for each topic that are then used as features to train a neural network classifier. In such studies, the specification of the information need is based on the topic of a document; much work has also been done on categorising documents by focusing on the stylistic aspect (e.g., genre classification and authorship attribution). For instance, Santini (2004) uses POS trigrams to categorise a subset of the BNC corpus into ten genres: four spoken genres (conversation, interview, public debate and planned speech) and six written genres (academic prose, advert, biography, instructional, popular lore and reportage). Stamatatos et al. (2000) (Automatic text categorisation in terms of genre and author) present an approach to text categorisation in terms of genre and author for Modern Greek. They extract a small set of style markers. We refer to Sebastiani (1999) for a substantial review on the machine learning models. There are also studies comparing among these machine learning models, including Yang and Liu (1999).

The Web brings new challenges to TC. Uprising application problems include spam and junk emails filtering and commercial Web page identification, etc. TREC-11 defines spam filtering tasks. Most classification tasks are on web pages, there are also requirements to classify Web sites; multi-layer (hierarchical) classification (taxonomy) is also one of the future directions.

6.7 QA

The success of the research on information retrieval in the last decade attracts attention to another related research area, that is, question answering. There has been a dramatic surge in interest in the study of natural language question answering since 1999, when the Question Answering track was first introduced in the Text Retrieval Conferences TREC-8³⁹. Different from IR, question answering allows users to query by questions in natural language (in contrast to a list of keywords in IR) and extract the exact answers (as opposed to retrieving relevant documents in IR).

Early work on question answering is described by Simmons (1965), who reviews more than 15 working systems. “These early systems include conversational question answerers, front-ends to structured data repositories and story understanding systems which try to find answers to questions from text sources” (Hirschman and Gaizauskas, 2001). Early dialogue systems (conversational question answerers) such as SHRDLU (Winograd, 1972) and GUS (Bobrow et al., 1977) were built as research systems to help researchers understand the issues involved in modelling human dialogue (Hirschman and Gaizauskas, 2001). In these systems, knowledge has been formally encoded in a knowledge base and therefore automatically extracting relevant information from raw texts is not an issue that such systems are concerned about. In the tradition of database-oriented systems, the question answering module provides a natural language interface for accessing the data stored in a database. Since there are standard languages for querying databases, the major work of the QA module is actually analysing questions and mapping a question into a standard database query. One representative system in this category is BASEBALL (Green et al., 1961, cited in Hirschman and Gaizauskas, 2001). Lehnert (1978) is a well-known work in story comprehension and question answering. She applies Schank and Abelson’s (1977) model of scripts and plans to model the psychological aspect in the process of question answering.

³⁹ TREC (Text REtrieval Conference). Refer to <http://trec.nist.gov/>

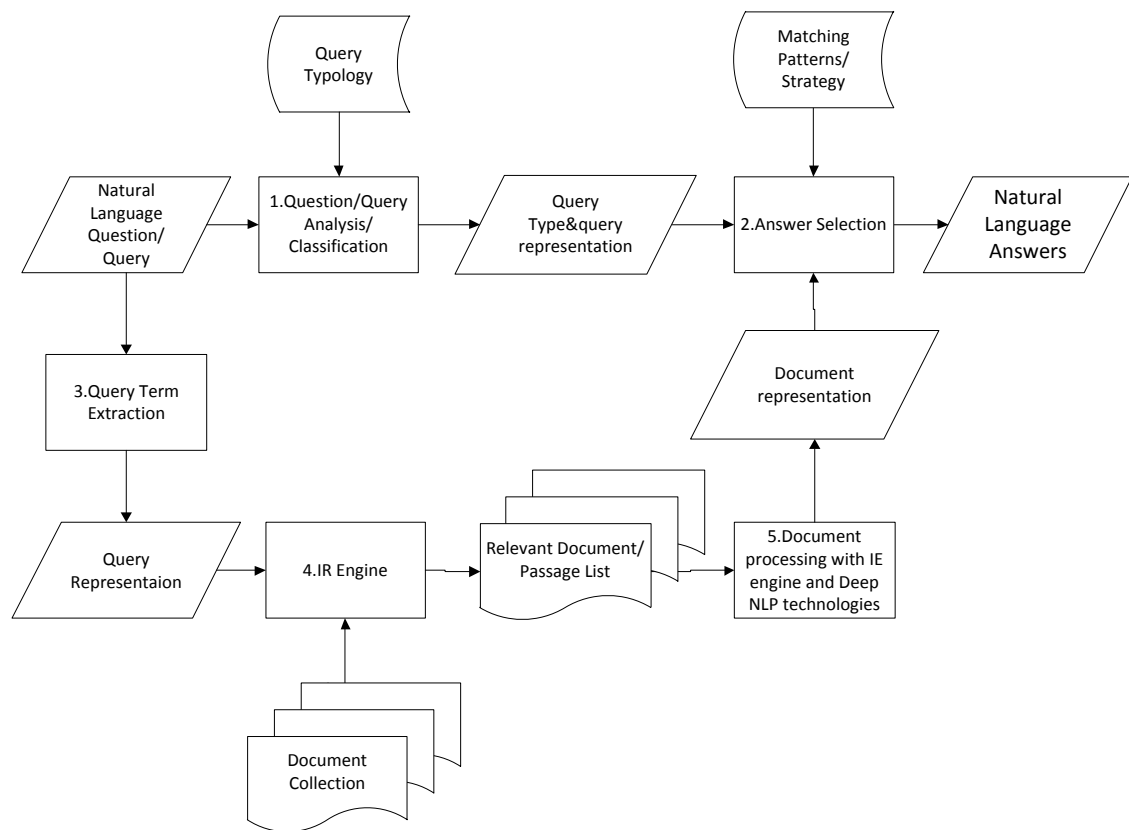


FIGURE 6-4. A GENERAL ARCHITECTURE OF QA SYSTEMS

All the above-mentioned systems are based on either hand crafted structured data or deep NLP techniques, both of which could only be applied to a limited domain. In recent years, the availability of large text collections both enables and necessitates the study of automatic answer extraction techniques for broader domains. Most of the current research on automatic question answering (QA) is driven by evaluation exercises such as TREC and CLEF⁴⁰, which focus on finding answers for factoid questions (questions that can be answered by a short phrase) from large text collections, such as Webclopedia (Hovy et al., 2001a) and Quetal (Neumann et al., 2003). There are also question answering systems that are based on the Web, such as START⁴¹ (Katz, 1997; Katz et al., 2006), Answer bus⁴² (Zheng, 2002) and LAMP (Zhang, 2004). IBM’s WATSON (Ferrucci, 2012) is a QA system that is specially designed to compete with human on the Jeopardy! quiz show. The Jeopardy! quiz covers broad categories including history, science and politics, etc. It is powered by the DeepQA project (Ferrucci et al.,

⁴⁰ CLEF (Cross Language Evaluation Forum). Refer to <http://clef.iei.pi.cnr.it/>

⁴¹ <http://start.csail.mit.edu/>

⁴² <http://www.answerbus.com/>

2010). The document collection is constructed using encyclopaedia, dictionaries and newswire articles, etc.

Current QA systems typically adopt the architecture shown in figure 6-4. We can see that for both the query and the document, there are two sets of processing operations. Specifically, query term extraction and query classification are applied to queries, and correspondingly, document term extraction (contained in the IR engine) and information extraction (IE) technology are applied to documents. Therefore, the matching process is divided into two stages: a low cost IR stage (module 3 and 4) which retrieves potentially relevant documents and a refined matching and answer extracting stage (module 2) based on deep textual processing results. We will introduce the major steps in section 6.7.1 to 6.7.3 respectively. As mentioned above, most of the systems focus on factoid questions, we will also introduce the direction of study non-factoid questions in section 6.7.4. There is one extra step in the DeepQA system. One important factor to be successful in the Jeopardy! quiz is to know how confident you are about your answer. Therefore the DeepQA system contains a module to score the confidence of the answers. One method is to attach the answer to the original question to form a hypothesis; they will then verify the hypothesis by searching it in the document collections and see how often it appears.

There are also systems that do not adopt the above-mentioned architecture. For example, the START system (Lin et al., 2002; Katz, 1997) uses a knowledge annotation approach. The raw content in Web documents are pre-annotated manually or semi-automatically with natural language sentences. When receiving a query, it matches the query to the annotation instead of the raw Web content. Both the annotation and the query are represented in T-expressions and they match against each other in the form of T-expressions. Since knowledge annotation is done manually or semi-automatically, it cannot be applied to the whole Web. It has been observed that the majority of the questions focus on a few types. For example, they find that users frequently asked about the demographics and economics of countries and the CIA world factbook⁴³ could provide answers to such questions. Therefore, the system can cover most of the questions by supporting a few resources such as the CIA world factbook.

⁴³ <https://www.cia.gov/library/publications/the-world-factbook/>

6.7.1 QUERY TERM EXTRACTION & INFORMATION RETRIEVAL ENGINE

Before retrieving relevant documents, a number of keywords are extracted from a query as its representation. Most QA systems extract content-bearing words from a question and use it to retrieve documents, such as in Webclopedia (Hovy et al., 2001a) and in Quetal (Neumann et al., 2003). For example, in Webclopedia, “Single- and multi-word units (content words) are extracted from the analysis, and WordNet synsets are used for query expansion. A Boolean query is formed.” (Hovy et al., 2001a; Hovy et al., 2001b)

Most systems use a standard search engine. For example, DeepQA uses the Indri⁴⁴ and the Lucene system; the KSAIL system (Katz et al., 2007) uses Lucene.

As we noted before, current QA systems mainly focus on answering factoid questions. The generic part of a factoid question is often encoded in the question word. For example, a question led by ‘where’ indicates the generic part is ‘name of a place’. So simply extracting content words leaves out the generic part of the question. For example, ‘how old is the Netherlands’ queen’ will be simplified as ‘Netherlands’ and ‘Queen’. We see that only focusing on the specific part in the first stage is similar to our proposed two-stage approach.

6.7.2 QUESTION CLASSIFICATION & IE

Given a question, the system will judge which conceptual class it actually addresses, this step is called question classification. After relevant documents or passages are retrieved, the system applies IE technology to analyse the texts and to assign semantic tags (conceptual class) to text snippets. Most QA systems stay at the level of simple named entity recognition (answering factoid questions). In recent several years, there is a trend of answering non-factoid questions, which will be introduced in section 6.8.4.

Most current QA systems only deal with a small number of question classes, such as PERSON NAME, LOCATION, TIME, NUMBER and DATE. Those classes are usually indicated by the question word. For example, the question word ‘how much’ indicates that the question is actually asking about a NUMBER. Sometimes, the question word itself is not enough for judging the semantic category; for instance, a question starting with ‘who’ might ask for a PERSON NAME, but also might ask *why a person is famous*. In Webclopedia (Hovy et al., 2001a), they use some linguistic features to disambiguate between these two categories.

⁴⁴ <http://ciir.cs.umass.edu/research/indri/>

In the above-mentioned question categories, some categories have regular internal structure, such as DATE, TIME and NUMBER (structured items); some are proper names, such as COMPANY NAME, LOCATION and PERSON NAME (names). Compared to structured items, names are much more difficult to extract. There are several reasons, including names are usually too numerous to be included in a dictionary or a Gazetteer, new names come out every day and “there are no strict rules governing the coinage of new names” (Appelt and Israel, 1999).

The cutting-edge techniques in named entity extraction model the internal structure of a named entity (if this named entity has regular structure) and the contextual features. Concrete methods can be divided into two strands: the knowledge engineering approach and the machine learning approach (Appelt and Israel, 1999). In the knowledge engineering approach, experts manually construct finite state rules based on the observation of a few annotated instances. The building blocks of such rules are typically characters, cue words, POS tags, etc. For example, a simple pattern for identifying a person’s name would be ‘Mr. <word>.’ One of the major drawbacks of the knowledge engineering approach is that the system cannot easily be adapted to new domains or languages. Machine learning approaches can still be divided into two categories, including the symbolic learning approach and the statistical based approach. Symbolic machine learning techniques induce a set of rules similar to those built manually by an expert. One unsupervised symbolic learning scenario is as follows: induce some rules based on a few annotated instances (seeds), use these rules to extract more instances and further induce more rules from the newly acquired instances; repeat this process until the extracted results do not change much in a new round. Typical statistical based models for Named Entity recognition are hidden Markov model, MEM and Connectionists models. Compared to symbolic approaches, statistical based machine learning models generally do not build explicit rules. Based on a large quantity of annotated data, the system can estimate the probability that a particular string belongs to a particular category. Detailed description of those models is outside the range of this thesis.

6.7.3 ANSWER SELECTION

After deep query and document analysis, the last step is to select the answer from a list of relevant documents or passages. This step can be considered as a matching process in which the similarity of the answer and the question are calculated. Systems differ with regards to the depth (in the sense of the level of text analysing) of the features they are using. One extreme

would be using the same approach as IR, in which the similarity between a question and a textual snippet is simplified as the similarity between two lists of keywords. Rather than pinpointing the answer, this method can only find some text snippets that contain potential answers. Most systems apply the result of named entity extraction and some surface lexical patterns for pinpointing answers. For example, for the question “when was X born”, the pattern for finding the answer would be “X was born in <DATE>” (Ravichandran and Hovy, 2002). Some systems match the semantic structure of a question and a potential answer by deep syntactic and semantic analysis. In Elworthy (2000), the system matches at the level of logic forms. In Leidner et al. (2004), a formal semantic representation model is defined for this purpose, based on Discourse Representation Theory.

Many systems combine the above approaches. In Weblopedia (Hovy et al., 2001a), they first perform pattern matching, then try to match the semantic category that a question demands; if all fails, they apply a simple word-matching algorithm.

6.7.4 ANSWERING NON-FACTOID QUESTIONS

In recent research on QA there is a trend towards answering non-factoid questions. In TREC, the complex interactive QA (ciQA) task started in 2006 (Dang et al., 2007; Dang et al., 2006). So far only a few types of non-factoid questions have been explored, including definitional, biographical and relationship questions (Katz et al., 2005; Lin, 2006). Among them, definitional questions (i.e., questions asking for the definition of something) are the most extensively studied. In TREC-10, Harabagiu et al. (2001) note that “the percentage of questions that ask for definitions of concepts represented 25% of the questions from the main task”, “an increase from a mere 9% in TREC-9 and 1% in TREC-8 respectively”.

The definition of a term usually contains a general conceptual category the term belongs to (i.e., the genus), and some important features differentiate it from similar things (the differentia). Within this general pattern, the details in a definition will vary from case to case. There are two major approaches to extracting the answers. The first is mining information from thesauruses or encyclopedias. Harabagiu et al. (2001) exploit relations of a term such as hypernym and synonym defined in wordnet. For example, they use hypernyms of a term to retrieve relevant sentences, based on the assumption that the hypernym is a kind of genus. Thus, for the question “what is shaman”, the hypernym of “shaman”, viz. “priest” or “non-Christian priest”, is used to retrieve the answers. Prager and Chu-Carroll (2001) adopt a similar approach, but with a more profound term-weighting schema. Another approach is based the

surface patterns or deep linguistic structure of a definition. In Blair-Goldensohn et al. (2003), they define syntactic and lexical patterns for some definitional predicates, such as genus and species. Xu et al. (2003) identify “kernel facts” for answering definitional questions by detecting appositive and copula constructions, identifying important predicate-arguments structures and handcrafting rules modelling typical definitional expressions.

6.8 SUMMARY

This chapter reviews various IR models and proposes that the current key word based IR approach does not fit for retrieving the generic part of extended topic. We present three reasons summarised as below.

- The generic part of an extended topic may not be explicitly expressed in a relevant document.
- A generic concept typically refers to a category of relationships, which are conveyed by cue phrases, discourse connectives or prepositions. Phrases become meaningless when being broken into a list of key words. Moreover, many cue phrases/discourse connectives/prepositions convey different relationships under different circumstances and are too ambiguous to retrieve good results when being separated out of the contexts.
- Directly apply a generic concept to match retrieving documents may bring noise since typical term weighting schemata such as the idf weighting schema are not able to accurately model the breadth of the contexts that generic concepts occur.

Query expansion could expand the original query to add more terms for retrieval and therefore partly solves the first problem. However, the effectiveness of this approach is largely confined by the first set of retrieved documents.

In section 6.4, we revisit the four classes of typical extended topic and propose different approaches for different classes. In chapter 7 and 8, we will use the proposed methodology to experiment on automatically retrieving relevant documents for several general question types, including causal, procedural and biographical. The proposed approaches have connections with several research areas, such as phrasal retrieval, NLP for IR, TC and QA. We also briefly introduce the literature on these areas in section 6.5 to 6.7.

CHAPTER 7

AUTOMATICALLY RETRIEVING RELEVANT DOCUMENTS FOR CAUSAL QUESTIONS

7.1 INTRODUCTION

This chapter addresses the problem of automatically retrieving documents to answer *causal questions*. Here by ‘causal question’ we refer to questions that inquire about the cause of a specific event, act, state or phenomenon. Automatically answering causal questions is important since our previous study on frequent asked questions in medical domain has shown that causal questions cover a larger proportion of questions than other eight question categories.

Causal questions consist of two parts: a) a specific part which refers to a specific event/act/state/phenomenon and b) a generic part – i.e., cause – which indicates the kind of required information about a). Although the cause of something could often be expressed in a very elaborated form, in which case, a network of events/propositions (related by complex relations) are included, however, here we will only focus on those simple cases, i.e., using one or a few sentences to summarise the cause. Such causal questions are representative of the second or the third question class described in section 4.2; the fourth question class, which is a more difficult case, will be addressed in the next chapter.

Following the general methodology proposed in chapter 6, we will use phrasal or linguistic patterns to detect causal relationship. In addition to the word ‘cause’, there are many other terms or phrases that signal a causal relationship, such as ‘is caused by’ and ‘arise from’. We call such terms or phrases as causal indicators. We collect causal indicators to expand the original query. It is worth pointing out that a causal relationship is not always explicitly expressed by a causal indicator. In such cases, a substantial amount of background knowledge is necessary to detect the implied causal relationship. There are studies (e.g., Bozsahin and Findler, 1992; Mooney, 1990) on reasoning inside a knowledge base to detect causality — i.e., the status of existing a causal relationship. However, these systems can hardly be enhanced to fit a different domain. Again, we stay at a shallow level, the aim being to design a domain-general method. Compared to typical information retrieval (IR) approach, which only uses the key words in a query to match relevant documents, the above-mentioned measure helps to discover more relevant documents and therefore improves the recall of the system.

Since causal indicators are used in a relatively broad context, directly using them to match documents may bring a lot of noise. To solve this problem and to increase the accuracy of the retrieving system, we invent a way to measure the likelihood of the association between the causal indicators and the inquired entity. The system only counts a causal indicator as a match if it is highly likely to be associated with the inquired entity. This is actually a question of ensuring the association between the specific part and the generic part of a question. In our study, we measure the strength of the association based on distances. To be more concrete, the idea is to combine each causal indicator and the inquired entity into a phrase, which has a predefined value representing the maximum allowed distance between the two elements; then use such phrase to match retrieve relevant documents. We call such phrase as linkage phrase.

This chapter is organised as follows: section 7.2 introduces a few related works; section 7.3 explains in details of the proposed method; section 7.4 explains how we collect linkage phrases; section 7.5 introduces the document scoring formula; section 7.6 and 7.7 talks about preparing the questions and the document collection for the retrieval experiment; section 7.8 provides the experiment results; we analyse the results in section 7.9; section 7.10 gives a short summary of the whole chapter.

7.2 RELATED WORK

Here we will review previous work on automatically detecting causal relationship and on phrase retrieval, each topic relates to a different aspect of our approach.

Both Girju (2003) and Khoo et al. (2000) provide a short review on automatically extracting causal relationship. They point out that previous systems developed in the 1980's or 1990's mostly rely on knowledge-based inferences to detect causal relationship. As knowledge is manually coded, such systems only focus on a very specific domain and can hardly be used for real applications. Recent studies apply simple linguistic patterns to identify causal relationship. For example, Khoo et al. (2000) uses a set of graphical patterns to match the syntactic tree of a sentence so as to identify causal relationship and to extract the *cause* and the *effect*. The key part of most of the patterns is a phrase which indicates causal relationship such as 'because of'. Similarly, Girju (2003) uses verbal expressions (e.g., 'lead to' and 'bring about') to detect causal relationship. Most of the collected expressions are ambiguous; therefore she devises a method to learn further semantic restrictions on the surrounding NPs. The semantic

restrictions are based on the category system defined in WordNet⁴⁵. She applied the collected patterns to a question answering system to extract answers. The experiment result shows that this approach performs much better than a default key-word-matching-based answer-extraction approach. While the above studies target specifically at detecting *causal* relationship, there are also many studies that target at a general method for detecting a predefined set of semantic or rhetorical relations between discourse units. These include Marcu (2000) and LeThanh et al. (2003), both of which largely rely on discourse connectives to detect semantic or rhetorical relations.

In the research of information retrieval, people sometimes use phrases rather than key words as the basic query units. This way they can ensure the key words in the document are related in the same way as in the question. The thinking of keeping the relation between keywords resembles our idea ensuring the relation between the causal indicator and the inquired entity. There are substantial studies on phrase retrieval. However, the conclusion seems not to be in agreement with each other: with some studies (Mitra et al., 1997; Carpenter, 2004) show that phrase retrieval does not outperform key word retrieval; but some others (Mishne and de Rijke, 2005; Croft et al., 1991) obtain the contrary.

7.3 METHODOLOGY

As mentioned before, we use linkage phrases to extend the original query to match retrieve relevant documents. A linkage phrase contains a causal indicator and a focused entity. For example, for question ‘What causes asthma?’, an example linkage phrase is ‘increase the risk of + asthma’. This approach resembles phrase matching in term of the thinking of keeping the association of the elements in the query.

As shown in Mishne and de Rijke (2005), exact phrase match does not perform as well as proximity phrase match. By proximity phrase match we mean that we allow the components of a phrase to be apart from each other as long as they are within a certain distance. We also define a maximum allowed distance for each linkage phrase. The linkage phrase ‘increase the risk of + asthma’ is redefined as ‘increase the risk of + <maximum allowed distance> + asthma’. We see that the redefined linkage phrase could match to sentence [7.1], in which there are two words in between ‘increase the risk of’ and ‘asthma’.

⁴⁵ Refer to <http://wordnet.princeton.edu/>

[7.1] A number of studies suggest that breast-feeding may increase the risk of eczema and asthma.

It is clear that the larger the maximum allowed distance is, the more matches we will find but the more noise it may also bring; and vice versa. In other words, there is a trade-off between accuracy and coverage. In our experiments, we compared between two approaches: a rigorous method that uses a relatively small distance value and a lenient method that uses a relatively large value (referred to hereafter as the first lenient method). Apart from increasing the distance, there is another way to relax the constraints of linkage phrase. The idea is that if the focused entity of a question contains two or more than two words, then instead of using one linkage phrase that contains all the key words, we use several linkage phrases, each of which contains only one key word. For example, if the question is ‘what is the cause of breast cancer’, then instead of using one linkage phrase which contains ‘breast cancer’ and a causal indicator such as ‘is caused by’, we use two linkage phrases: the first one is ‘breast + <maximum allowed distance> + is caused by’ and the second one is ‘cancer +<maximum allowed distance> + is caused by’. We refer to this method as the second lenient method.

7.4 CAUSAL INDICATORS PREPARATION

We gathered causal indicators from several resources, including the causal verbs provided by Girju (2003), the causal expressions by Khoo et al. (2000) and the discourse connectives provided by Marcu (2000), LeThanh (2003) and Oates (2001). Many of these phrases are highly ambiguous. For example, the verb ‘produce’ is often used in the same sense with ‘manufacture’; only occasionally it is used to denote causal relationship. We searched each indicator in the medical collection described in section 7.7 and only used those frequent and not very ambiguous ones. Finally we got 35 causal indicators. The causal indicators and the usage examples are listed in appendix F.1.

As mentioned before, we evaluate the likelihood of the association between a causal indicator and the inquired entity based on the distance between them. Therefore, we defined a maximum allowed distance for each causal indicator. Apart from that, the inquired entity must be in the right direction (i.e., either before the causal indicator or after it) to be something being caused. For example, for the causal indicator ‘be caused by’, we expect the thing being caused to be before it. We specified a pattern for each causal indicator to represent these requirements. These patterns will be matched against texts to identify useful features. Table 7-1 shows the patterns specified for all the causal indicators in group 1 and

group 2. Refer to appendix F.2 for patterns of other groups. Since an IR system indexes word stems rather than the original word, therefore, the defined patterns contain only word stems. The symbol '+' in the patterns simply means concatenation, which does not match anything. Here *causee* refers to the thing being caused.

As mentioned before, the maximum allowed distance varies across different causal indicators. The value is acquired by observing real-life usage of these phrases. Specifically, we grouped causal indicators into ten groups. Presumably the causal indicators inside one group have similar usage. We collected a set of expressions from the web for each group, the size of which varies from 30 to 126. We marked the position of the causee as well as the causal indicator and extract the distance between them. We used different maximum allowed distance for different retrieval methods: the largest number in each set was applied in the first lenient method (the second number in each cell in column 'Max Value'); after removing the repetitive numbers in each set, the median of the remaining numbers was used in the rigorous method and the second lenient method (the first number in each cell in column 'Max Value').

Group	Causal Phrase	Pattern	Max Value
1	'caused by'	<causee>(+<wordstem>){0..Max}+caus+by	10/52
	'induced by'	<causee>(+<wordstem>){0..Max}+induc+by	
	'provoked by'	<causee>(+<wordstem>){0..Max}+provok+by	
	'evoked by'	<causee>(+<wordstem>){0..Max}+evok+by	
	'triggered by'	<causee>(+<wordstem>){0..Max}+trigger+by	
	'stimulated by'	<causee>(+<wordstem>){0..Max}+stimul+by	
2	'cause'	caus(+<wordstem>){0..Max}+<causee>	4/23
	'induce'	induc(+<wordstem>){0..Max}+<causee>	
	'provoke'	provok (+<wordstem>){0..Max}+<causee>	
	'evoke'	evok(+<wordstem>){0..Max}+<causee>	
	'trigger'	trigger(+<wordstem>){0..Max}+<causee>	

TABLE 7-1. PATTERNS OF LINKAGE PHRASE

7.5 SCORING SCHEMA

We adapted the vector space model to score documents. Each querying term or phrase is defined as one dimension in the vector space and the score of a document is defined as the similarity between the query vector and the document vector. The value of each dimension in a vector is proportional to the frequency of the term/phrase (i.e., tf) and its reversed document frequency (i.e., idf). There are different ways to calculate the similarity between two vectors — i.e., $\text{sim}(q,d)$. We used the vector space model delivered in the Lucene-1.4.2 package⁴⁶, in which $\text{sim}(q,d)$ is defined as the inner product of the two vectors. Refer to formula 7.1 for the precise definition.

$$\text{sim}(q, d) = \sum_{t \in q} \frac{tf_{t,q} \cdot idf_t}{norm_q} \cdot \frac{tf_{t,d} \cdot idf_t}{norm_d} \cdot coord_{q,d} \quad (7.1)$$

Where

$$tf_{t,X} = \sqrt{\text{freq}(t, X)} \quad (7.2)$$

$$idf_t = 1 + \log \frac{|D|}{(\text{freq}(t, D) + 1)} \quad (7.3)$$

$$coord_{q,d} = \frac{|q \cap d|}{|q|} \quad (7.4)$$

$$norm_q = \sqrt{\sum_{t \in q} (tf_{t,q} \cdot idf_t)^2} \quad (7.5)$$

$$norm_d = \sqrt{|d|} \quad (7.6)$$

$|d|$ and $|q|$ represent the length of a document d and the length of a query q respectively. $q \cap d$ represents the set of terms in common between the query and the document. $|D|$ represents the size of the document set D and $\text{freq}(t, D)$ refers to the number of documents that contains term t .

⁴⁶ Refer to <http://lucene.apache.org/>

Since distance is an indicator of the likelihood of the association, for a particular match to the linkage phrase, the longer the distance, the lower it should be scored. The above way to define tf does not reflect this requirement. Besides, since documents are indexed as a term list in an IR system, it is also expensive to calculate the idf of a phrase. We kept the above definition of tf and idf for single term but redefined them for phrases. The new definition of the similarity between a query and a document ($sim'(q, d)$) is as follows.

$$sim'(q, d) = \sum_{t \in q} \frac{tf_{t,q} \cdot idf_t}{norm_q} \cdot \frac{tf_{t,d} \cdot idf_t}{norm_d} + \sum_p \frac{tf_{p,d} \cdot idf_p^2}{norm_q \cdot norm_d} \quad (7.7)$$

where p refers to a linkage phrase; idf_p is calculated as in 7.8; $tf_{p,d}$ is calculated as in 7.9.

$$idf_p = \sum_{t \in p} idf_t \quad (7.8)$$

$$tf_{p,d} = \sqrt{sloppyfreq(p, d)} \quad (7.9)$$

The sloppy frequency of a phrase p in a document d is the sum of the score assigned to each occurrence pi in the document. Formula 7.10 is used to calculate this score.

$$sloppyfreq(p, d) = \sum_{pi \in d} \frac{1}{\left(\frac{len_{pi} - len_p}{mdis_p} + 1 \right)} \quad (7.10)$$

Here len_p refers to the minimum length of the linkage phrase; $mdis_p$ refers to the median distance in the collected examples of the linkage phrase, see section 7.4 for the definition. We can see that $tf_{p,d}$ decreases when len_{pi} increases.

$norm_q$ in 7.7 is as defined in 7.11.

$$norm_q = \sqrt{\sum_{t \in q} (tf_{t,q} \cdot idf_t)^2 + \sum_p (idf_p)^2} \quad (7.11)$$

7.6 QUESTION PREPARATION

Medicinenet⁴⁷ archives over 6000 medical questions and their answers. MayoClinic⁴⁸ also provides a list of questions asked to physicians. We went through these two question

⁴⁷ <http://www.medicinenet.com/>

collections to select a small set of causal questions. We only considered those questions that contain one or two key words in their specific parts. To ensure there are enough relevant documents in our document collection (described in section 7.7), we searched the specific part of each candidate question and removed those questions with few matches (i.e., less than 10). Through the above process, we finally obtained 21 questions, ten of which contain only one key word in their specific parts, the rest contain two key words, as listed in table 7-2.

Questions
What causes asthma?
What causes joint pain?
What causes diabetes?
What causes gastrointestinal bleeding?
What causes glaucoma?
What causes male infertility?
What causes constipation?
What causes Crohn's disease?
What causes angina?
What causes atrial fibrillation?
What causes stroke?
What causes heart failure?
What causes epilepsy?
What causes lung cancer?
What causes melanoma?
What causes myocarditis?
What causes white piedra?
What causes mesothelioma?

⁴⁸ <http://www.mayoclinic.com/health/AnswersIndex/AnswersIndex>

What causes tinea?
What causes neuropathic pain?
What causes pulmonary hypertension?

TABLE 7-2. QUESTION COLLECTION

7.7 CORPUS PREPARATION AND RELEVANCE ASSESSMENT

The corpus that we experimented with is provided by the MUCHMORE project⁴⁹. This corpus contains 9668 medical scientific abstracts initially obtained from the Springer web site⁵⁰. “Abstracts are taken from 41 medical journals (e.g. *Der Nervenarzt*, *Der Radiology*, etc.), each of which constitutes a homogeneous medical sub-domain (e.g. Neurology, Radiology, etc.)” (Volk et al., 2003). This corpus has a wide coverage. This can be reflected by the fact that we searched 91 disease names (has a broad topic coverage) in the collection and only four of them had no match. However, some brief examination also indicated that it contains only a few relevant documents for the questions in our prepared question collection. We therefore searched each question in Medline⁵¹ and downloaded the top ranked ten abstracts. 210 abstracts were downloaded and added into the corpus. Example documents from the Muchmore project and from Medline are included in appendix F.3.

Relevance judgement is done by the author of this thesis. For each question, we searched it in the corpus using different ranking methods; then the top 20 results from different ranking methods were extracted and merged together. Each question and document pair was marked as either ‘relevant’ or ‘irrelevant’.

We had many difficulties in assessing document relevance. One major problem is the difficult medical terminology. For example, it is necessary for us to know that pulmonary hypertension is not a specific kind of hypertension to avoid mistaking documents relevant to the cause of the hypertension as relevant to the cause of pulmonary hypertension.

There are many ambiguous cases as well. For example, the passage below talks about the cause of postoperative intestinal atonia; we may also induce, from the first sentence, that postoperative intestinal atonia occurs together with constipation. In such case, it is difficult to

⁴⁹ <http://muchmore.dfki.de/>

⁵⁰ <http://link.springer.de/>

⁵¹ <http://www.ncbi.nlm.nih.gov/entrez/query.fcgi>

say whether the above passage is also about the cause of constipation. In our experiment, we generally take ambiguous cases as relevant.

[7.1] Postoperative intestinal atonia is a complication which is likely to occur in patients predisposed for constipation and in patients after intra-abdominal operations. The postoperative delay of bowel movement, however, is often also related to the type of anaesthesia being used. In order to evaluate the magnitude of an anaesthetic-induced postoperative delay of bowel movement, two types of intravenous-based anaesthesia using fentanyl/midazolam (1 mg/25 mg; dosage 0.1 ml/kg/h), and ketamine/midazolam (250 mg/25 mg; dosage 0.1 ml/kg/h) respectively were compared with a volatile anaesthetic technique (enflurane; mean concentration 1.5 vol%). ...

Below are some other principles we have applied in the relevance assessment.

a. a passage is relevant even if there is only one phrase that is talking about the cause of the inquired disease. For example, the following passage is considered relevant to ‘the cause of constipation’ since the phrase ‘morphine-induced constipation’ provides an answer.

[7.2] “Almost all patients treated with opioids suffer from constipation. Numerous laxatives are used to overcome the problem, but none has yet been found to yield favourable results in all patients. Several studies have attempted to reverse opioid-induced constipation by the use of oral naloxone. Experiments carried out in rats showed that morphine-induced constipation is reduced by oral naloxone without impairment of antinociception. ...”

b. a passage is relevant if instead of providing the cause of a disease, it rules out some causing factor of a disease, such as the passage below.

[7.3] industrial pollutants and agricultural pesticides do not cause breast cancer, a survey shows. ...

c. a passage is relevant if it is talking about the cause of a disease that is a more general or a more specific problem than the inquired disease. For example, cancer is a more general problem than breast cancer.

d. a passage is irrelevant if it mentions the topic without providing an answer. For example, the following passage is not relevant to the question asking about ‘the cause of inflammatory bowel diseases’, even though it contains the phrase.

[7.4] ... despite continued uncertainty about the cause of inflammatory bowel diseases, recent advances nourish the hope for further improvement of the control of disease activity and a better quality of life for patients with inflammatory bowel diseases. ...

7.8 EXPERIMENTS AND RESULTS

The baseline results are acquired by typical key-word-matching-based IR methods. The generic part as well as the specific part is applied to retrieve relevant documents. We experimented on two scoring methods, i.e., the vector space model implemented in the Lucene-1.4.2 package and the BM25 scoring schema implemented in the Terrier IR platform⁵².

We compared the rigorous method and the first lenient method against the baseline. The average precisions of using different methods are shown in table 7-3.

	precision@5	precision@10	precision@20
Vector Space	0.514286	0.47619	0.390476
BM25	0.542857	0.466667	0.37381
Rigorous Method	0.790476	0.633333	0.469048
Lenient Method I	0.790476	0.638095	0.485714

TABLE 7-3. AVERAGE PRECISIONS OF FOUR DIFFERENT METHODS USING ALL THE QUESTIONS

The best result in each category is highlighted in table 7-3. The results show that both the rigorous method and the first lenient method perform much better than the two baseline methods. We also did a pairwise t-test to test the significance of the differences between the methods, as shown in table 7-4. Each cell in this table contains three numbers, representing the significance of differences between two methods at precision@5, precision@10 and precision@20 respectively.

⁵² Refer to <http://ir.dcs.gla.ac.uk/terrier/>

	Vector Space	BM25	Rigorous Method	Lenient Method I
Vector Space	-	.267/.505/.308	.000/.000/.000	.000/.000/.001
BM25	-	-	.000/.000/.001	.000/.000/.000
Rigorous Method	-	-	-	1.00/.803/.167
Lenient Method I	-	-	-	-

TABLE 7-4. THE SIGNIFICANCE OF DIFFERENCES AMONG FOUR METHODS USING ALL 21 QUESTIONS

We see from the above table that the difference between the baseline and our proposed two methods are significant. There is no significant difference between the Vector Space model and BM25, neither is there significant difference between the rigorous method and the first lenient method.

To evaluate the second lenient method, we recalculated the above numbers for only the ten questions that have two key words in its specific part. The result is shown in table 7-5.

	precision@5	precision@10	precision@20
Vector Space	0.52	0.52	0.425
BM25	0.52	0.51	0.42
Rigorous Method	0.84	0.67	0.525
Lenient Method I	0.82	0.68	0.55
Lenient Method II	0.7	0.6	0.49

TABLE 7-5. AVERAGE PRECISIONS OF FIVE DIFFERENT METHODS USING PART OF THE QUESTIONS

The best result is again achieved by either the rigorous method or the first lenient method, as highlighted. The performances of these two systems are pretty close, which are far better than the two baseline systems. The result of the second lenient method is about in the middle,

slightly closer to the first two proposed methods. Table 7-6 gives the significance of the differences using pairwise t-test.

	Vector Space	BM25	Rigorous Method	Lenient Method I	Lenient Method II
Vector Space	-	1.00/.758/.823	.000/.013/.012	.000/.001/.002	.041/.193/.089
BM25	-	-	.001/.018/.012	.002/.010/.001	.095/.247/.039
Rigorous Method	-	-	-	.591/.758/.052	.089/.173/.271
Lenient Method I	-	-	-	-	.193/.104/.051
Lenient Method II	-	-	-	-	-

TABLE 7-6. THE SIGNIFICANCE OF DIFFERENCES AMONG FIVE METHODS USING PART OF THE QUESTIONS

The above table shows the same pattern with table 7-4 when only considering the first four methods. It also shows that there is no significant difference between the third proposed method (lenient method II) and the first two. When compared to the baseline methods, the third proposed method performs significantly better than the Vector Space model and BM25 at precision@5 and precision@20 respectively.

7.9 DISCUSSION

The experiment results have shown that both the rigorous method and the first lenient method significantly outperform classical key-word-matching based IR method. The average precisions of the second lenient method are also larger than those of the baseline systems. However, pairwise t-test shows that many of the differences are not statistically significant.

We also observe that the proposed methods improve the baseline more at higher positions. For example, in table 7-3, the first lenient method improves the basic vector space model by 53.7% at precision@5, by 34.0% at precision@10 and by 24.6% at precision@20. We would say this is due to that in our proposed ranking methods we apply a parameter measuring the strength of the association between the causal indicators and the inquired entity. This parameter, compared to the key-word-frequency-based parameters, greatly increases the accuracy of the prediction.

7.10 SUMMARY

This chapter investigates methods for automatically retrieving documents to answer causal questions. We suggest to using two measures to enhance a typical key-word-based IR system: one is to use a list of causal indicators to extend the original query so that more relevant documents could be identified; another is to use the distance between the causal indicators and the inquired entity to measure the strength of association between them. We design three methods, one rigorous method and two lenient methods. All of the three methods incorporate the above two measures and only differ from each other in some detailed configuration. The result shows that the rigorous method and the first lenient method significantly outperform the baseline and gain particular better result at higher positions. Causal question is representative of a set of extended topic (i.e., the second question class defined in section 4.2). The above proposed method represents a general solution we propose for this set of extended topics. As one of the future directions, we can apply the proposed method to other question types.

CHAPTER 8

AUTOMATICALLY RETRIEVING RELEVANT DOCUMENTS FOR PROCEDURAL AND BIOGRAPHICAL QUESTIONS

8.1 INTRODUCTION

This chapter addresses the problem of automatically retrieving relevant documents for procedural and biographical questions. By procedural questions, we refer to questions inquiring about the *procedure* for achieving a specific goal, e.g., ‘how to make an omelette?’ or ‘how to install Windows XP server?’. Procedural questions constitute a large proportion of how-to questions and are therefore a popular question type. By biographical questions, we refer to questions that ask for biographical key facts such as profession, birthdates and education, which are typically found in biographies, e.g., ‘Who was Martin Luther King Jr.?’ or ‘Who was Virginia Woolf?’. It is noted (Voorhees, 2003) that definitional questions occur relatively frequently in logs of web search engines. A lot of definitional questions are asking about the biography of a person. Different from causal questions, which often retrieve one or two sentences, both of these two types of questions usually retrieve one to several passages containing a network of related facts/events/propositions. This is a shared trait of the fourth question class described in section 4.2. We experiment on these two types of questions to explore a general solution for the fourth question class.

We again view each type of questions consisting of two parts: a specific part and a generic part. For procedural questions, the specific part would be the content of the goal (e.g., install Windows XP server) and the generic part would be procedural (i.e., a series of steps for achieving this goal). For biographical questions, the specific part is usually a person’s name (e.g., George Clooney) and the generic part specifies the nature of the required information about the person (i.e., biographical facts). Typical information retrieval (IR) methods, based on key word matching, are better suited to detecting the specific part than the generic part.

As previously noted, the generic part is often implicit. To resolve this problem, we use a list of phrasal/linguistic patterns to detect key procedural/biographical elements. This is similar to what was done in the last chapter to detect causes. For example, we use cue phrase ‘first’, ‘next’ and ‘then’ to detect sequential relationships between actions, which are key elements of a procedure. One single sequential relationship would not suffice to say the whole document is about a procedure; therefore, we train a classifier to score document based on

Chapter 8 – Automatically Retrieving Relevant Documents for Procedural and Biographical Questions

the proportion of the procedural text they contain. By ‘procedural text’ we refer to ordered lists of steps, which are very common in some instructional genres such as online manuals. We follow the same solution to detect biography.

The other problem is how to combine the two scores of a document, representing, respectively, how relevant it is to the specific part and to the generic part. As previously noted, a general concept occurs in a much broader context than a specific entity. In other words, a document relevant to the specific entity is more likely to be relevant than a document fitting a general concept. In this case, the generic part may bring a lot of noise when applied together with the specific part to retrieve documents. For causal questions, we suggested to use linkage phrases to combine causal indicator and the specific topic instead of using causal indicators separately. This approach cannot be applied to biographical or procedural questions. This is because, different from a cause, a procedure or a biography is usually a long stretch of texts, including many detailed facts or relationships, which are not necessarily directly relates to the specific entity. We instead suggest a two stage approach to solve this type of questions: (1) use typical IR approaches to retrieve documents that are relevant to the specific part; (2) use a text categoriser to only re-rank the retrieved documents based on how relevant they are to the generic part (biography or procedural in this case). As text categorisation approach to detecting the generic part is very expensive, the two-stage approach also makes it feasible to apply as in step 2 only a small set of documents are processed. As noted in section 6.3, query expansion technique is able to automatically mine related keywords to expand the original query. In this chapter, we will compare our method against a query expansion method.

This chapter is organised as follows: in section 8.2, we introduce our experiments on procedural questions; section 8.3 introduces our experiments on biographical questions; section 8.4 gives a short summary.

8.2 AUTOMATIC RETRIEVING PROCEDURES

8.2.1 INTRODUCTION

This section is organised as follows. Section 8.2.2 introduces some related work on answering procedural questions. Section 8.2.3 talks in detail about how we score document *procedurality* -- i.e., the proportion of procedural text they contain. In particular, we will introduce a new classifier adapted from a Naïve Bayes Classifier to better fit our less than ideal training corpus. Section 8.2.4 presents the experiments on automatically retrieving relevant documents for

Chapter 8 – Automatically Retrieving Relevant Documents for Procedural and Biographical Questions

procedural questions. We compare among three approaches, including only use the specific part to retrieve document, our two-stage approach and a query expansion approach. Section 8.2.5 discusses the experiment results.

8.2.2 RELATED WORK

Only a few studies have addressed procedural questions. Murdok and Croft (2002) distinguishes between “task-oriented questions” (i.e., ask about a process) and “fact-oriented questions” (i.e., ask about a fact) and presents a method to automatically classify questions into these two categories. Following this work, Kelly et al. (2002) explore the difference between documents that contain relevant information to the two different types of questions. They conclude, “lists and FAQs occur in more documents judged relevant to task-oriented questions than those judged relevant to fact-oriented questions” (Kelly et al., 2002: 645) and suggest, “retrieval techniques specific to each type of question should be considered” (Kelly et al., 2002: 647). Schwitter et al. (2004) presents a method to extract answers from technical documentations for How-questions. To identify answers, they match the logical form of a sentence against that of the question and also explore the typographical conventions in technical domains. The work that most resembles ours is Takechi et al. (2003), which uses word n-grams to classify (as procedural or non-procedural) list passages extracted using HTML tags. Our approach, however, applies to a whole document and uses more complicated phrasal/linguistic features.

8.2.3 RANKING PROCEDURAL TEXTS

Three essential elements of a text categorisation approach are the features used to represent the document, the training corpus and the machine learning method, which will be described in section 8.2.3.1, 8.2.3.2 and 8.2.3.3 respectively. Section 8.2.3.4 presents experiments on applying the learned model to rank documents.

8.2.3.1 FEATURE SELECTION AND DOCUMENT REPRESENTATION

LINGUISTIC FEATURES AND CUE PHRASES

We targeted six procedural elements: actions, times, sequences, conditionals, preconditions and purposes. These elements can be recognised using linguistic features or cue phrases. For example, an action is often conveyed by an imperative; a precondition can be expressed by the cue phrase ‘only if’. We used all the syntactic and morphological tags defined in

Chapter 8 – Automatically Retrieving Relevant Documents for Procedural and Biographical Questions

Connexor’s syntax analyser⁵³. There are some redundant tags in this set. For example, both the syntactic tag ‘@INFMARK>’ and the morphological tag ‘INFMARK>’ refer to the infinitive marker ‘to’ and therefore always occur together at the same time. We calculated the Pearson’s product-moment correlation coefficient (r) (Weisstein, 1999) between any two tags based on their occurrences in sentences of the whole training set. We removed one in each pair of strongly correlated tags and finally got 34 syntactic tags and 34 morphological tags.

We also handcrafted a list of relevant cue phrases (44), which were extracted from documents by using the Flex tool for pattern matching. Some sample cue phrases and the matching patterns are shown in table 8-1. Table G-1 in appendix G.1 includes all the cue phrases and syntactic and morphological tags.

Procedural Element	Cue Phrase	Pattern
Precondition	‘only if’	[Oo]nly[:space:]if[:space:]
Purpose	‘so that’	[sS]o[:space:]that[:space:]
Condition	‘as long as’	(([Aa]s) [:space:]long[:space:]as[:space:])
Sequence	‘first’	[fF]irst [:space:][:punct:]
Time	‘now’	[nN]ow[:space:][:punct:]

TABLE 8-1. SAMPLE CUE PHRASES AND MATCHING PATTERNS

MODELLING INTER-SENTENTIAL FEATURE COOCCURRENCE

Some cue phrases are ambiguous and therefore cannot reliably suggest a procedural element. For example, the cue phrase ‘first’ can be used to represent a ranking order, a spatial relationship as well as a sequential order. However, it is more likely to represent a sequential order among actions if there is also an imperative in the same sentence. Indeed, sentences that contain both an ordinal number and an imperative are very frequent in procedural texts. We compared between the procedural training set and the non-procedural training set to extract distinctive feature cooccurrence patterns (limited to 2 features). Two schemas were used.

Chi-square was applied to measure the significance of the correlation between whether a sentence contains a particular feature cooccurrence pattern and whether it is in a procedural

⁵³ <http://www.connexor.com/>

Chapter 8 – Automatically Retrieving Relevant Documents for Procedural and Biographical Questions

document. Table 8-2 is the contingency table for a cooccurrence pattern, where p_{pro} , n_{pro} , p_{non} and n_{non} stand for the total number of sentences of each category.

	In procedural documents	In non-procedural documents
Contain the pattern	p_{pro}	p_{non}
Do not contain the pattern	n_{pro}	n_{non}

TABLE 8-2. THE 2*2 CONTINGENCY TABLE FOR A FEATURE COOCCURRENCE PATTERN

Chi-square favours patterns that occur frequently. There are patterns that do not occur in every procedural document, but if they do occur, there is a high probability that the document is procedural. Such a pattern might have a low frequency of occurrence in the corpus and therefore cannot generate a significant chi-square value. To detect such patterns, another schema we used is the ratio between the number of sentences that contain a particular pattern in the procedural set (p_{pro}) and in the non-procedural set (p_{non}), normalised by the size of the two sets (s_{pro} and s_{non}), as shown in formula 8.1.

$$R(p) = \frac{p_{pro} \times s_{non}}{p_{non} \times s_{pro}} \quad (8.1)$$

Two ordered lists were acquired by applying the two schemas to rank the feature cooccurrence patterns. We cut the two lists at certain thresholds (which were decided empirically) and acquired two sets of top ranked patterns. Those patterns that were included in both sets were chosen as distinctive patterns.

DOCUMENT REPRESENTATION

Each document was represented as a vector $\vec{d}_j = \{x_{1j}, x_{2j}, \dots, x_{Nj}\}$, where x_{ij} represents the number of sentences in the document that contains a particular feature normalised by the document length. We compare the effectiveness of using individual features (x_{ij} refers to either a single linguistic feature or a cue phrases) and of using feature co-occurrence patterns (x_{ij} refers to a feature co-occurrence pattern).

8.2.3.2 CORPUS PREPARATION

Pagewise⁵⁴ provides a list of subject-matter domains, ranging from household issues to arts and entertainment. We downloaded 1536 documents from this website (referred to hereafter as the Pagewise collection). We then used some simple heuristics to select documents from this set to build the initial training corpus. Specifically, to build the procedural set we chose documents with titles containing key phrases ‘how to’ and ‘how can I’ (209 documents); to build the non-procedural set, we chose documents which did not include these phrases in their titles, and which also had no phrases like ‘procedure’ and ‘recipe’ within the body of the text (208 documents). Appendix G.2 provides a sample document from Pagewise.

Samples drawn randomly from the procedural set (25) and the non-procedural set (28) were submitted to two judges, who assigned procedural scores from 1 (meaning no procedural text at all) to 5 (meaning over 90% procedural text). The Kendall tau-b agreement between the two rankings is 0.821. Overall, the average scores for the procedural and non-procedural samples were 3.15 and 1.38. We used these 53 sample documents as part of the test set and the document remaining as the initial training set (184 procedural and 180 non-procedural).

This initial training corpus is far from ideal. First, it is small in size. In our experiments, we used this initial training set to bootstrap a larger training set. Details will be described in section 8.2.3.4. A more serious problem is that many positive training examples do not contain a major proportion of procedural texts. To mitigate this problem, we designed an adapted Naive Bayes classifier, details will be introduced in section 8.2.3.3.

8.2.3.3 LEARNING METHOD

As mentioned in previous sections, we adapted the Naive Bayes classifier to better fit our suboptimal training corpus. Before describing the details of the adaptation, we first talk about another problem we came across while experimenting with the Naive Bayes classifier.

We used the Naive Bayes classifier from the Weka-3-4 package (Witten and Frank, 2000). Some preliminary experiments showed that most documents were scored as either extremely procedural (i.e., the score is 1) or not procedural at all (i.e., the score is 0). Such scoring result does not enable us to rank the documents. We analysed and modified the Naive Bayes classifier to solve the problem. Details are described as follows.

⁵⁴ Refer to <http://www.essortment.com>

*Chapter 8 – Automatically Retrieving Relevant
Documents for Procedural and Biographical Questions*

The Naive Bayes classifier scores the degree of procedurality using the probability that a document falls into the procedural category—i.e., $p(C = \textit{procedural} | \vec{d}_j)$. Using the Bayes' theorem, the probability can be calculated as shown in

$$p(C = c | \vec{d}_j) = \frac{p(C = c)p(\vec{d}_j | C = c)}{p(\vec{d}_j)} = \frac{p(C = c)p(\vec{d}_j | C = c)}{p(C = c)p(\vec{d}_j | C = c) + p(C = -c)p(\vec{d}_j | C = -c)} \quad (8.2)$$

where c can represent any particular category (e.g., procedural)⁵⁵.

Assuming that any two coordinates in the document feature vector $\vec{d} = \{x_1, x_2, \dots, x_N\}$ are conditionally independent, we can then simplify the calculation by using

$$p(\vec{d}_j | C = c) = \prod_i p(X_i = x_i | C = c) \quad (8.3)$$

where $p(X_i = x_i | C = c)$ represents the probability of randomly picking up a document in category c of which the feature X_i has value x_i . The same simplification applies to $p(\vec{d}_j | C = -c)$.

Multiplying all the $p(X_i = x_i | C = c)$ together often yields an extremely small value that is difficult to represent in a computer; this is why the final procedurality score is either 1 or 0. To tackle this problem, we calculated the procedurality score by

$$\log\left(\frac{p(C = c | \vec{d}_j)}{p(C = -c | \vec{d}_j)}\right) = \log\left(\frac{p(C = c)p(\vec{d}_j | C = c)}{p(C = -c)p(\vec{d}_j | C = -c)}\right) \quad (8.4)$$

$$= \log(p(C = c)) - \log(p(C = -c)) + \sum_i \log(p(X_i = x_i | C = c)) - \sum_i \log(p(X_i = x_i | C = -c)) \quad (8.5)$$

The above modification will be referred to hereafter as the first adaptation. It is worth pointing out that the ranking order of any two documents remains the same when replacing (2) by (5).

As we can see from the above formulas, the Naive Bayes classifier scores a document according to whether it is a typical member of its set (i.e., $p(X_i = x_i | C = c)$) as well as how much it contrasts with members of other set (i.e., $\frac{1}{p(X_i = x_i | C = -c)}$). Specifically, the Naive

⁵⁵ Note that formula (8.2), (8.3) and (8.6) are either extracted from (John and Langley, 1995) or inferred from the Java code in the Weka-3-4 package.

Chapter 8 – Automatically Retrieving Relevant Documents for Procedural and Biographical Questions

Bayes classifier delivered in the Weka-3-4 package assumes that each feature follows a normal distribution and estimates $p(X_i = x_i | C = c)$ by

$$p(X_i = x_i | C = c) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_i - \mu_i)^2}{2\sigma_i^2}} \quad (8.6)$$

where μ and σ are estimated from the training data. In figure 8-1, the solid curve and the dotted curve show the probability density functions estimated from the non-procedural training set and the procedural training set respectively.

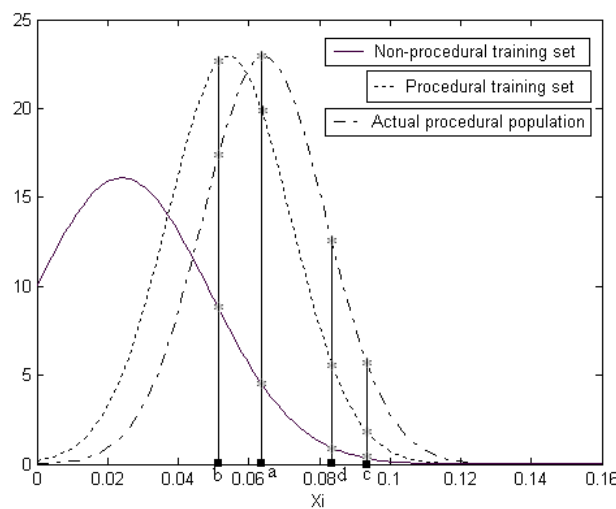


FIGURE 8-1. AN ILLUSTRATION OF THE PROBLEMS IN USING THE NAIVE BAYES CLASSIFICATION ALGORITHM

As mentioned in section 8.2.3.2, the procedural documents in our initial training corpus have a low average procedurality score, which means many of them do not contain a large proportion of procedural elements. Since most of the features used represent some procedural elements, we suppose the actual population of procedural documents should have a higher mean value on such features compared to the procedural training set (as shown in figure 8-1). In this case, point a, which obviously has a higher score than point b when using a training set that are representative of the actual population, is probably scored lower than b when using our training set.

Although the procedural training examples are not representative of the actual population of procedural documents, they are useful in indicating the difference between procedural documents and non-procedural documents. For example, we can infer from figure 8-1 that feature X_i is associated positively with the degree of procedurality (since the positive training set has a higher mean value on this feature than the negative training set). However, the

Chapter 8 – Automatically Retrieving Relevant Documents for Procedural and Biographical Questions

Naive Bayes classifier does not focus on modelling the difference between the two different classes. Therefore, point c, although larger than point d, is probably scored lower than d. We adjusted the formula to model the difference between the two different classes. Specifically, we replaced $p(X_i = x_i | C = c)$ in formula 8.6 by

$$\left\{ \begin{array}{ll} p(X_i \leq x_i | C = c) = \int_{-\infty}^{x_i} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx & \text{if } \text{mean}(X_i | C = c) > \text{mean}(X_i | C = -c) \\ p(X_i \geq x_i | C = c) = \int_{x_i}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx & \text{if } \text{mean}(X_i | C = c) < \text{mean}(X_i | C = -c) \\ 1 & \text{if } \text{mean}(X_i | C = c) = \text{mean}(X_i | C = -c) \end{array} \right. \quad (8.7)$$

where $\text{mean}(X_i | C = c)$ refers to the mean value of feature X_i of the documents in category c . $p(X_i = x_i | C = -c)$ in (5) was replaced by a formula similar to (7), the only difference being that every c in (7) is changed to be $-c$.

The new scoring curves are shown in figure 8-2. This way the score of a document is determined by the ratio of the probability of a document with a lower feature value being in the procedural class (represented by dotted curve) and the probability of a document with a higher feature value being in the non-procedural class (represented by the solid curve). If feature X_i is associated negatively with the degree of procedurality (i.e., $\text{mean}(X_i | C = c) < \text{mean}(X_i | C = -c)$), then the word ‘lower’ and ‘higher’ in the last sentence should be reversed.

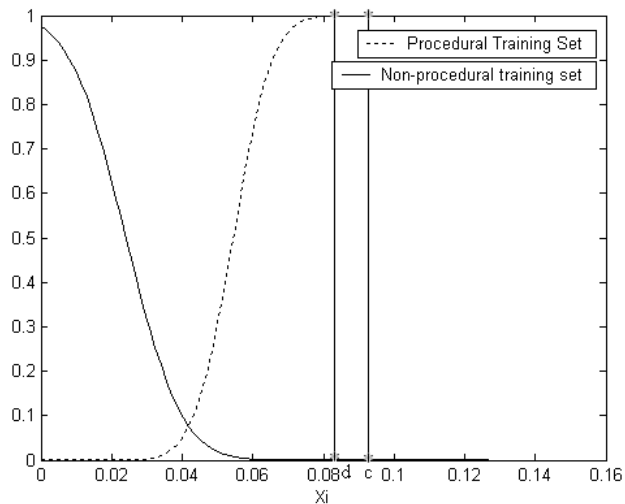


FIGURE 8-2. AN ILLUSTRATION OF THE ADAPTED NAIVE BAYES CLASSIFICATION ALGORITHM

The new model will be referred to hereafter as the Adapted Naive Bayes classifier. It is worth pointing out that, after the above modification, (5) is no longer equivalent to (4). This means,

*Chapter 8 – Automatically Retrieving Relevant
Documents for Procedural and Biographical Questions*

the core of the Naive Bayes classification approach — i.e., viewing the “categorization status value” in terms of the probability that a document falls within a category (Sebastiani, 1999) is changed.

8.2.3.4 EXPERIMENTS

Our training and testing corpora were from two sources: the Pagewise collection and the SPIRIT collection. The SPIRIT collection contains a terabyte of HTML that are crawled from the Web starting from an initial seed set of a few thousands universities and other educational organisations (Clarke et al., 1998). Appendix G.2 provides sample document from Pagewise and SPIRIT.

Our test set contained 103 documents, including the 53 documents that were sampled and then separated from the initial training corpus, another 30 documents randomly chosen from the Pagewise collection and 20 documents chosen from the SPIRIT collection. We asked two human subjects to score the procedurality for these documents, following the same instruction described in section 8.2.3.2. The correlation coefficient (Kendall tau-b) between the two rankings is 0.725, which is the upper bound of the performance of the classifiers.

As mentioned before, the initial training corpus was used to bootstrap a larger training set. To do so, we first extracted 441 distinctive feature cooccurrence patterns based on the initial training corpus. These patterns were used to build document vectors to train an Adapted Naive Bayes classifier. We applied the classifier to rank the remaining documents from the Pagewise collection (the whole set excluding 83 documents that were added into the test set) and 500 web documents from the SPIRIT collection. 378 top ranked documents were selected to construct the positive training set and 608 lowest ranked documents were used to construct the negative training set. A random sampling of the procedural documents in this bootstrapped set suggests that their average procedurality score is slightly higher than those in the initial training set.

The bootstrapped training corpus was then used to reselect distinctive feature cooccurrence patterns and to train different classifiers. We compared the Adapted Naive Bayes classifier with the Naive Bayes classifier⁵⁶ and three other classifiers, including Maximum Entropy

⁵⁶ As mentioned in section 8.2.3.3, we cannot rank the documents based on the scoring result of the Naive Bayes classifier from the Weka-3-4 package. We therefore used the model acquired after the first adaptation instead.

Chapter 8 – Automatically Retrieving Relevant Documents for Procedural and Biographical Questions

(ME)⁵⁷, Alternating Decision Tree (ADTree) (Freund and Mason, 1999) and Linear Regression (Witten and Frank, 2000).

Figure 8-3 and table 8-3 show the kendall-tau b coefficients between human subjects' ranking results and the trained classifiers' ranking results on the test set when using individual features (112). Figure 8-4 and table 8-4 show the kendall-tau b coefficients when using feature cooccurrence patterns (813).

The figures show that when using individual features, Linear Regression achieved the best result, Adapted Naive Bayes performed the worst, Naive Bayes, Maximum Entropy and Alternating Decision Tree were in the middle; when using feature cooccurrence patterns, the order almost reversed, i.e., Adapted Naive Bayes performed the best and Linear Regression the worst. Comparing the results of using individual features and feature cooccurrence patterns, only Adapted Naive Bayes and Naive Bayes performed better when using feature cooccurrence patterns, all the other classifiers performed better using individual features.

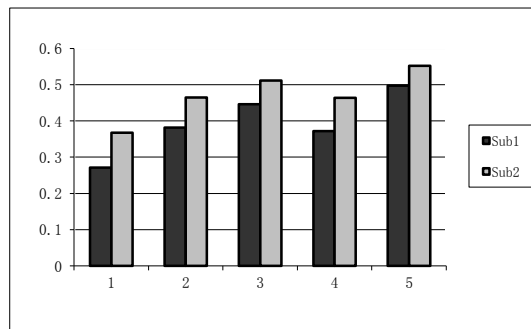


FIGURE 8-3. RANKING RESULTS USING INDIVIDUAL FEATURES: 1 REFERS TO ADAPTED NAIVE BAYES, 2 REFERS TO NAIVE BAYES, 3 REFERS TO ME, 4 REFERS TO ADTREE AND 5 REFERS TO LINEAR REGRESSION

Ranking Method	Agreement With Subject1	Agreement with Subject 2	Average
Adapted Naive Bayes	0.270841	0.367515	0.319178
Naive Bayes	0.381921	0.464577	0.423249
Maximum Entropy	0.446283	0.510926	0.478605
Alternating Decision Tree	0.371988	0.463966	0.417977
Linear Regression	0.497395	0.551597	0.524496

TABLE 8-3. RANKING RESULTS USING INDIVIDUAL FEATURES

⁵⁷ Refer to <http://homepages.inf.ed.ac.uk/s0450736/maxent.html>

Chapter 8 – Automatically Retrieving Relevant Documents for Procedural and Biographical Questions

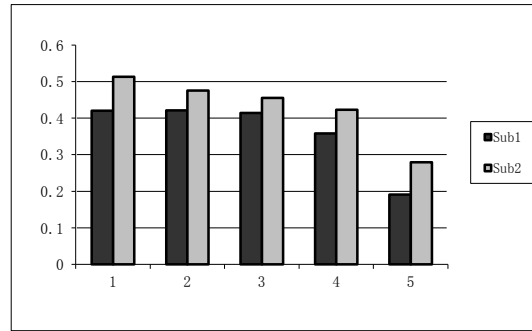


FIGURE 8-4. RANKING RESULTS USING FEATURE COOCCURRENCE PATTERNS: 1 REFERS TO ADAPTED NAIVE BAYES, 2 REFERS TO NAIVE BAYES, 3 REFERS TO ME, 4 REFERS TO ADTREE AND 5 REFERS TO LINEAR REGRESSION

Ranking Method	Agreement with Subject 1	Agreement with Subject 2	Average
Adapted Naive Bayes	0.420423	0.513336	0.466880
Naive Bayes	0.420866	0.475514	0.448190
Maximum Entropy	0.414184	0.455482	0.434833
Alternating Decision Tree	0.358095	0.422987	0.390541
Linear Regression	0.190609	0.279472	0.235041

TABLE 8-4. RANKING RESULTS USING FEATURE COOCCURRENCE PATTERNS

8.2.3.5 DISCUSSION

The experiment results showed that two Naive Bayes classifiers fit better with feature cooccurrence patterns while ME, ADTree and Linear Regression fit better with individual features. In contrast to feature cooccurrence patterns, each of which is chosen as being very distinctive, individual features may contain many irrelevant features since all the morphological and syntactical taggers that the Connexor’s Syntax Analyser provides are included. This does not make much difference for ADTree and Linear Regression since they both have a feature selection process that can filter irrelevant features. However, the Adapted Naive Bayes classifier does not have such a function and it treats every feature as extremely distinctive. The Naive Bayes classifier and the ME classification model, although do not have an explicit feature selection process, can estimate the degree of distinctiveness of each feature based on the training data. The above difference between the Naive Bayes classifier and the Adapted Naive Bayes classifier can be seen by comparing the scoring results at point d in figure 8-1 (Naive Bayes classifier) and figure 8-2 (Adapted Naive Bayes classifier).

To verify the above explanation with regard to why the Adapted Naive Bayes classifier performed poorly when using individual features, we applied the feature selection algorithms

Chapter 8 – Automatically Retrieving Relevant Documents for Procedural and Biographical Questions

described in section 8.2.3.1 to select distinctive individual features (42 features were selected⁵⁸) and tested the classifiers again. Results are shown in figure 8-5 and table 8-5.

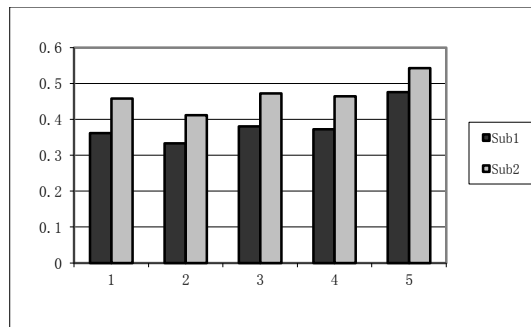


FIGURE 8-5. RANKING RESULTS USING SELECTED INDIVIDUAL FEATURES: 1 REFERS TO ADAPTED NAIVE BAYES, 2 REFERS TO NAIVE BAYES, 3 REFERS TO ME, 4 REFERS TO ADTREE AND 5 REFERS TO LINEAR REGRESSION

Ranking method	Agreement with Subject 1	Agreement with Subject 2	Average
Adapted Naive Bayes	0.362007	0.458198	0.410103
Naive Bayes	0.332798	0.411717	0.372258
Maximum Entropy	0.380151	0.471868	0.426010
Alternating Decision Tree	0.371988	0.463966	0.417977
Linear Regression	0.476054	0.542832	0.509443

TABLE 8-5. RANKING RESULTS USING SELECTED INDIVIDUAL FEATURES

Compared to using all the individual features, the performance of Adapted Naive Bayes was greatly improved when only using a few selected ones; ADTree performed the same; but ME and Linear Regression performed slightly worse. This does support our thinking that the Adapted Naive Bayes classifier presumes every feature is extremely distinctive and therefore only distinctive features should be used with it.

Another important difference between feature cooccurrence patterns and individual features consists in their numbers. Because the number of feature cooccurrence patterns is huge, it is difficult for ADTree and Linear Regression to generalise and to select relevant features in such a high feature vector space. This can be a reason why these two models performed poorly when using feature cooccurrence patterns.

⁵⁸ Table G.2 in appendix G includes all the features in this set.

Chapter 8 – Automatically Retrieving Relevant Documents for Procedural and Biographical Questions

As it was addressed in section 8.2.3.1, individual features are ambiguous and we expected modelling feature inter-sentential cooccurrence helps in disambiguate. However, this thinking is not supported by the results of the experiments. We believe this is because that most documents used in the experiments are from the Pagewise collection, which have a rather uniform style and ambiguities are greatly decreased.

8.2.4 RETRIEVING RELEVANT DOCUMENTS FOR HOW-TO QUESTIONS

In this section we will describe the experiments on retrieving relevant documents for how-to questions by applying different approaches mentioned in the introduction section 8.2.1.

8.2.4.1 EXPERIMENT SETUP

We randomly chose 60 how-to questions from the query logs of the FAQ finder system (Burke et al., 1997). Three judges went through these questions and agreed on 10 *procedural questions*⁵⁹.

Questions
How do I cook a herring?
How do I set up PPP in LINUX?
How do you build a fire?
How do I format a CDROM?
How to build a bicycle?
How do I set up anonymous FTP in UNIX?
How do I start a home based business?
How to install windows NT?
How do I install a car stereo?
How to lose weight?

TABLE 8-6. PROCEDURAL QUESTION SET

We searched in Google and downloaded 40 top ranked documents for each question, which were then mixed with 1000 web documents randomly sampled from the SPIRIT collection to compile a test set. The reason we used 40 documents is to make sure there are enough relevant results for question. One sample document from Google is included in appendix G.2. The two-stage architecture is as shown in figure 8-6. In the first stage, we sent only the

⁵⁹ We distinguish questions asking for a series of steps (i.e., procedural questions) from those of which the answer could be a list of useful hints, e.g., ‘how to make money’.

Chapter 8 – Automatically Retrieving Relevant Documents for Procedural and Biographical Questions

content of the goal to a state-of-the-art IR model to retrieve 30 documents from the test set, which were reranked in the second stage according to the degree of procedurality by a trained document classifier.

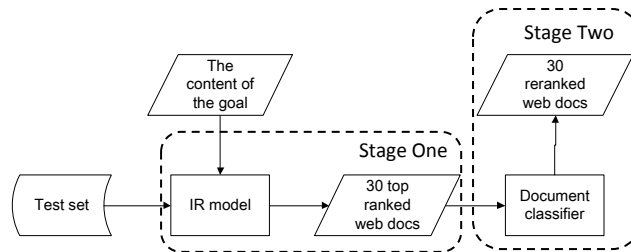


FIGURE 8-6. A TWO-STAGE ARCHITECTURE

We also tried to test how well query expansion could help in retrieving procedural documents, following a process as shown in figure 8-7. First, key words in the content of the goal were used to query an IR model to retrieve an initial set of relevant documents, those of which that do not contain the phrase ‘how to’ were then removed. The top 10 remaining documents were used to generate 40 search terms, which were applied in the second round to retrieve documents. The reason we used 10 documents is because we observed that the top ten documents contained a large proportion of relevant documents and there were a lot of irrelevant documents further down the list. We chose to use 40 search terms for query expansion is because rough examination indicated that 40 top ranked terms were relevant to the topic but not beyond that. Finally the 30 top ranked documents were returned as relevant documents.

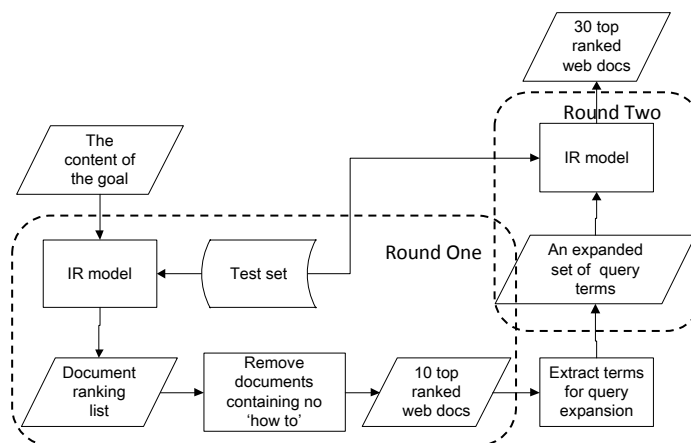


FIGURE 8-7. AN ALTERNATIVE ARCHITECTURE USING QUERY EXPANSION

8.2.4.2 IR MODEL

For the above-mentioned IR model, we used the BM25 and PL2 algorithms from the Terrier IR platform⁶⁰.

The BM25 algorithm is one variety of the probabilistic schema presented in (Robertson et al. 1993). It has gained much success in TREC competitions and has been adopted by many other TREC participants.

The PL2 algorithm, as most other IR models implemented in the Terrier IR platform, is based on the Divergence From Randomness (DFR) framework. Amati and van Rijsbergen (2002) provide a detailed explanation of this framework and a set of term-weighting formulae derived by applying different models of randomness and different ways to normalise the weight of a term according to the document length and according to a notion called *information gain*. They test these different formulae in the experiments on retrieving relevant documents for various sets of TREC topics and show that they achieve comparable result with the BM25 algorithm.

We also used the Bo1 algorithm from the same package to select terms for query expansion. Refer to (Plachouras et al., 2004) for details about this algorithm.

8.2.4.3 RESULT

We tested eight systems, which could be organised into two sets. The first set uses BM25 algorithm as the basic IR model and the second set uses PL2 as the basic IR model. Each set includes four systems: a baseline system that returns the result of the first stage in the two-stage architecture, one system that uses query expansion technique following the architecture in figure 8-7 and two systems that apply the two-stage architecture (one uses the Adapted Naive Bayes classifier and another one uses the Linear Regression classification model).

The mean average precision (MAP)⁶¹ of different retrieval systems is shown in table 8-7 and figure 8-8.

⁶⁰ <http://ir.dcs.gla.ac.uk/terrier/index.html>

⁶¹ The average precision of a single question is the mean of the precision scores after each relevant document is retrieved. The mean average precision is the mean of the average precisions of a collection of questions.

Chapter 8 – Automatically Retrieving Relevant Documents for Procedural and Biographical Questions

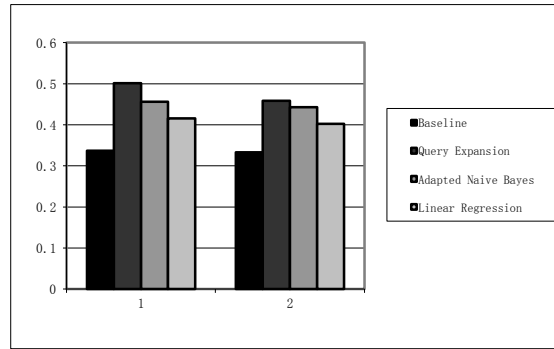


FIGURE 8-8. MAPS OF DIFFERENT SYSTEMS: 1 REFERS TO USING BM25 AS THE IR MODEL, 2 REFERS TO USING PL2 AS THE IR MODEL.

	Model	MAP
Set1	BM25 (Baseline)	0.33692
	BM25 + Query Expansion	0.50162
	BM25 + Adapted Naive Bayes	0.45605
	BM25 + Linear Regression	0.41597
Set2	PL2 (Baseline)	0.33265
	PL2 + Query Expansion	0.45821
	PL2 + Adapted Naive Bayes	0.44263
	PL2 + Linear Regression	0.40218

TABLE 8-7. RESULTS OF DIFFERENT SYSTEMS.

We can see that in both sets: (1) systems that adopt the two-stage architecture performed better than the baseline system but worse than the system that applies query expansion technique; (2) the system that uses Adapted Naive Bayes classifier in the second stage gained better result than the one that uses Linear Regression classification model. We performed a pairwise t-test to test the significance of the difference between the results of the two systems with an integrated Adapted Naive Bayes classifier and of the two baseline systems. Each data set contained 20 figures, with each figure representing the average precision of the retrieving result for one question. The difference is significant ($p=0.02$). We also performed a pairwise t-test to test the significance of the difference between the two systems with an integrated Adapted Naive Bayes classifier and of the two systems using query expansion techniques. The difference is not significant ($p=0.66$).

8.2.5 DISCUSSION

Contrary to our expectation, the results of the experiments showed that the two-stage approach did not perform better than simply applying a query expansion technique to

*Chapter 8 – Automatically Retrieving Relevant
Documents for Procedural and Biographical Questions*

generate an expanded list of querying terms. We provide the following explanation for this result.

First, query expansion is able to automatically expand the original query with a list of related terms which signals detailed procedural elements. Secondly, the detailed design of the retrieval process using query expansion technique already embodies the thinking of extended topic. Specifically, in the first stage, to acquire an initial list of relevant documents, we do not use the generic part (i.e., ‘how to’ or ‘procedure’) to directly match retrieve relevant documents, which may bring a lot of noise; instead we first use the specific part to retrieve a ranked list of documents and then filter those that do not contain ‘how to’. In other words, we do not use ‘how to’ as a ranking condition but use it as a filter condition. Besides, we had thought that many documents that contain procedures do not contain the word ‘procedure’ or the phrase ‘how to’; however, we found that such words or phrases, although not included in the body of the text, often occur in the title of the document.

8.3 AUTOMATIC RETRIEVING BIOGRAPHIES

8.3.1 INTRODUCTION

This section focuses on automatically retrieve relevant documents for biographical questions. We compare among four approaches: : a. only use the specific part to retrieve documents; b. use the specific part and the generic part together to retrieve documents; c. apply the query expansion technique; d. the aforementioned two-stage approach.

This section is organised as follows: section 8.3.2 provides a short literature review; section 8.3.3 presents in details how we train classifiers to automatically identify biographical documents, which are to be integrated into the two-stage approach; section 8.3.4 introduces the implementation of the four approaches and reports the results of the retrieval experiments; Section 8.3.5 gives a short summary.

8.3.2 RELATED WORK

The work is relevant to three research areas, including Text Categorisation (TC), Information Retrieval (IR) and Question Answering (QA).

Mainstream research in TC focuses on categorising a few standard corpora, such as the Reuter 21578 corpus (Lewis, 1997), in which documents are organised into a list of pre-defined categories. The category of ‘biographical facts’ has never been applied.

*Chapter 8 – Automatically Retrieving Relevant
Documents for Procedural and Biographical Questions*

The TREC competition⁶² started in 1992 and has been held once per year since then. It provides a platform where systems can compare against each other and might be the most important annual event in recent IR studies. Queries about biographical information rarely occurred in early TREC competitions. We have seen biographical questions in the web track (retrieving web documents) and in the filtering track (essentially study of text categorisation); however, there is no work in these two tracks targeted specifically for such questions.

In 1998, the question answering (QA) track was first introduced into the TREC competition. As mentioned in section 6.7.4, most work in QA focuses on factoid questions and it was not until 2003 that significant participation in answering definitional questions started (biographical question is considered as a kind of definition question). It is noted (Voorhees, 2003) that definitional questions occur relatively frequently in logs of web search engines and are an important type of question. The only work that we are aware of on retrieving documents to answer biographical questions is by Tsur et al. (2004). They experiment with a subset of biographical questions in the QA track in TREC 2003, the aim being to test how well external sources (in comparison to a few standard corpora) can help answering these questions. To do so, they first search a question in Google and acquire a set of potentially relevant documents; then use a classifier to identify biographical documents from this set; finally extract answers from these documents. One problem in common between our experiments and theirs is training a text classifier to identify biographical documents.

8.3.3 TEXT CATEGORISATION FOR IDENTIFYING BIOGRAPHICAL DOCUMENTS.

8.3.3.1 CORPUS PREPARATION

TRAINING SET

To compile the training set, we downloaded 1255 biographies from [biography.com](http://www.biography.com)⁶³. Most biographies in this set are short and have a uniform style. To prevent the trained classifiers being geared towards the particular style of this set, we also downloaded 100 biographies from several other biography indices on the Web such as [answers.com](http://www.answers.com)⁶⁴ and [astrutriths.com](http://www.astrutriths.com)⁶⁵,

⁶² Refer to <http://trec.nist.gov/>.

⁶³ Refer to <http://www.biography.com>

⁶⁴ Refer to <http://www.answers.com>

⁶⁵ Refer to <http://www.astrutriths.com>

Chapter 8 – Automatically Retrieving Relevant Documents for Procedural and Biographical Questions

etc. The negative training set includes 500 documents extracted from the Reuters-21578 corpus and 676 English webpages randomly sampled from the SPIRIT collection. An example document from the SPIRIT collection is provided in appendix G.2; examples documents from other sources are provided in appendix H.1.

TEST SET

The test set contains 188 documents in total, 100 of which are biographies and 88 of which are non-biographies. The biography set includes 40 biographies from biography.com, 10 from several other biography indices on the Web and 50 documents collected by querying Google for biographies of a few persons. The non-biography set includes 50 documents from the Reuter-21578 collection and 33 English webpages from the SPIRIT collection.

8.3.3.2 DOCUMENT PREPROCESSING AND FEATURE SELECTION

DOCUMENT PREPROCESSING

The biography of a person usually contains his/her birthday, profession, deeds or contribution to the society, big moves in his/her life, etc. There are stereotypical phrases (pattern) to convey the above facts. For example, the pattern '<PERSON NAME> was born in <DATE>' is often used to convey a person's birthday; 'he/she worked as a/an <PROFESSION>' can be used to convey a person's profession, etc. We design ways to automatically extract such patterns.

<YEAR>	Four digits surrounded by white space.
<MONTH>	The twelve months from 'January' to 'December' or their abbreviations such as 'Jan.' and 'Feb.'.
<DATE>	A set of patterns such as '<MONTH> {1 or 2 digits}, <YEAR>'
<PROFESSION>	Compiled a set of 879 professions
<SINGLE_THIRD_PERSON_PRONOUN>	Combines <FEMALE_PRONOUN> and <MALE_PRONOUN>

TABLE 8-8. FIVE META-TAGS

We first used LingPipe to preprocess the documents. LingPipe is a suite of Java tools designed to perform linguistic analysis on natural language data. The particular tool we used is a statistical named-entity detector, which is able to mark up PERSON, LOCATION, ORGANISATION, FEMALE_PRONOUN (i.e., 'she' and 'her') and MALE_PRONOUN (i.e., 'he' and 'his'). We also designed a simple text annotation system, which further marks up YEAR,

Chapter 8 – Automatically Retrieving Relevant Documents for Procedural and Biographical Questions

MONTH, DATE, PROFESSION and SINGLE_THIRD_PERSON_PRONOUN. These meta-tags are explained in table 8-8.

FEATURE SELECTION

Once named entities had been marked up in the training corpus, we counted the frequency of every single n-gram (n is between 1 and 6) and kept those that occur at least in 30 sentences in the 1355 biographical documents in the training set. Note that any single punctuation, word or meta-tag is considered as 1-gram. We got more than one thousand of patterns. We design several feature selection algorithms to reduce this set.

We first calculated the Phi coefficient⁶⁶ to measure the degree of association between the occurrence of a pattern in a sentence and whether the sentence is in a biography. The contingency table is shown in table 8-9, where p_{bio} , n_{bio} , p_{non} and n_{non} stand for the total number of sentences of each category. The higher the Phi coefficient is, the more distinctive the pattern is. Patterns were ordered into a list according to the Phi coefficient.

	In biographical documents	In non-biographical documents
Contain the pattern	p_{bio}	p_{non}
Do not contain the pattern	n_{bio}	n_{non}

TABLE 8-9. THE 2*2 CONTINGENCY TABLE FOR MEASURING THE DISTINCTIVENESS OF A PATTERN

Some of the patterns are strongly correlated. Two patterns are strongly correlated means that if one pattern occurs in a sentence, then mostly probably the other will also occur in the same sentence and vice versa. An example is pattern ‘was born’ and pattern ‘born in’. We removed one pattern from any two strongly correlated patterns. Specifically, we calculated the Pearson product-moment correlation coefficient (r) between any two patterns based on their occurrence in the whole training set (46969 sentences in total); we then extracted pairs of patterns that has a strong correlation ($r \geq 0.88$) and removed the one in each pair that have a smaller Phi coefficient.

⁶⁶ Refer to http://en.wikipedia.org/wiki/Phi_coefficient

Chapter 8 – Automatically Retrieving Relevant Documents for Procedural and Biographical Questions

We cut the list at the top 400 pattern and the top 200 pattern respectively; therefore acquired two pattern sets (referred to hereafter as featureset1 and featureset2). Both of these two pattern sets will be used in the text categorisation experiment.

	In biographical documents	In non-biographical Documents
Contain pattern1 and pattern2	p_{bio}	p_{non}
Contain pattern 1 but not pattern2	n_{bio}	n_{non}

TABLE 8-10. THE 2*2 CONTINGENCY TABLE FOR CALCULATING PHI2

	In biographical documents	In non-biographical documents
Contain pattern2 but not pattern 1	p_{bio}	p_{non}
Do not contain pattern 2	n_{bio}	n_{non}

TABLE 8-11. THE 2*2 CONTINGENCY TABLE FOR CALCULATING PHI3

Observing featureset1 and featureset2, we found that there are pairs of patterns with one pattern being a substring of the other one. For example, ‘studied at’ is a substring of ‘studied at an <ORGANISATION>’. It might be redundant to include both of the two patterns into the feature set, since often only one of them is truly distinctive. We designed algorithms to judge whether to keep both of them or to remove one. We definitely keep the one that has a relatively high Phi coefficient (referred to as pattern1) and use it as a base to judge whether to remove the other one (referred to as pattern2). Specifically, if pattern1 is a substring of pattern2, then we calculated the Phi coefficient that measures the degree of association between whether a sentence contains pattern2 given the fact that it contains pattern1 and whether the sentence is in a biographical document. The contingency table is shown in table 8-10. To differentiate this Phi coefficient from the one mentioned before, we refer to it as Phi2. If pattern2 is a substring of pattern1, then we calculated the Phi coefficient that measures the degree of association between whether a sentence contains pattern2 but not pattern1 and whether the sentence is in a biographical document (referred to as Phi3). The contingency table is shown in table 8-11. Only if the above-mentioned association is significant ($\text{Phi2} > 0.04$ or $\text{Phi3} > 0.04$) do we keep pattern2.

Chapter 8 – Automatically Retrieving Relevant Documents for Procedural and Biographical Questions

The algorithm described in the last paragraph was applied to a subset of highly distinctive patterns ($\Phi > 0.04$). We got 278 patterns that constitute featureset3. Appendix H.2 provides a subset of all the patterns in featureset3.

8.3.3.3 DOCUMENT REPRESENTATION

Each document is represented as a feature vector $\vec{d} = \{x_1, x_2, \dots, x_N\}$, where x_i represents the number of sentences containing a particular pattern in the document (pf), normalised by the document length l . We adopted the formula used in Amati and van Rijsbergen (2002) to normalise a feature weight according to the length of a document, as shown in $x_i = pf \cdot \log_2 \left(1 + \frac{avg_l}{l} \right)$, where avg_l stands for the average document length of the whole training set.

8.3.3.4 CLASSIFIERS

We applied the Support Vector Machine (SVM), the Naive Bayes Classifier and the Alternating Decision Tree (ADTree) from the Weka-3-4 package (Witten and Frank, 2000). Since the adapted Naive Bayes Classifier is designed to tackle the problem with the training corpus in the experiment on classifying procedural texts, we did not apply it here.

8.3.3.5 EXPERIMENTS AND RESULTS

We used our training set to train different classification models which were then applied to classify the documents in the test set. The results are shown in figure 8-9 and table 8-12, where accuracy is defined as the ratio between the number of samples that are correctly classified and the total number of samples.

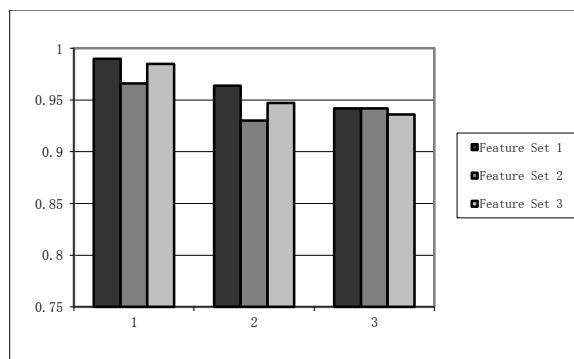


FIGURE 8-9. TEXT CATEGORISATION RESULT: 1 REFERS TO NAIVE BAYES, 2 REFERS TO SVM, 3 REFERS TO ADTREE

Chapter 8 – Automatically Retrieving Relevant Documents for Procedural and Biographical Questions

Classifier	Feature Set	Accuracy
Naive Bayes	Set 1	0.989
	Set 2	0.962
	Set 3	0.984
SVM	Set 1	0.962
	Set 2	0.929
	Set 3	0.945
ADTree	Set 1	0.94
	Set 2	0.94
	Set 3	0.934

TABLE 8-12. TEXT CATEGORISATION RESULT

The average performance of the Naive Bayes classifier when using different feature sets is the best of the three classifiers. Comparing the result of using different feature sets, the order is featureset1 > featureset3 > featureset2 (except with ADTree). The above results are considerably better than the results reported in Tsur et al. (2004), which are 89% for a Ripper-based algorithm and 83% for SVM. We applied the two best classification models (i.e., Naive Bayes with featureset1 and Naive Bayes with featureset3) for the document retrieval experiments

8.3.4 AUTOMATICALLY RETRIEVING DOCUMENTS FOR BIOGRAPHICAL QUESTIONS

8.3.4.1 QUESTIONS PREPARATION

We used the same set of biographical questions used in Tsur et al. (2004). As mentioned in section 8.3.2, they aim to extract the answers for the questions while we aim to just retrieve relevant documents, therefore, our results cannot be compared to theirs directly. However, an answer string extraction module can be easily plugged into our system in the future.

Questions

Who is Alberto Tomba?

Who is Albert Ghiorso?

Who is Alexander Pope?

Chapter 8 – Automatically Retrieving Relevant Documents for Procedural and Biographical Questions

Who is Alice Rivlin?
Who is Absalom?
Who is Nostradamus?
Who is Machiavelli?
Who is Andrea Boccelli?
Who is Al Sharpton?
Who is Aga Khan?
Who is Ben Hur?

TABLE 8-13. BIOGRAPHICAL QUESTION SET

The set of questions is shown in table 8-13. We removed the last question (i.e., who is Ben Hur) from this set since Ben Hur is a fictional character.

8.3.4.2 SYSTEM ARCHITECTURE

As mentioned in section 8.3.1, we experimented with four approaches: the first two just use a typical IR model, with one only using the specific part of a question to query the IR model and another using both the specific part and the generic part; the third one—i.e., applying query expansion technique—is shown in figure 8-10; the fourth one—i.e., the two-stage architecture—is shown in figure 8-11.

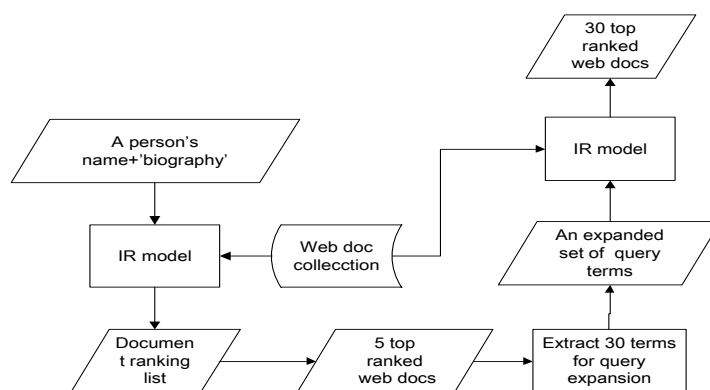


FIGURE 8-10. THE SYSTEM THAT APPLIES THE QUERY EXPANSION TECHNIQUE

Chapter 8 – Automatically Retrieving Relevant Documents for Procedural and Biographical Questions

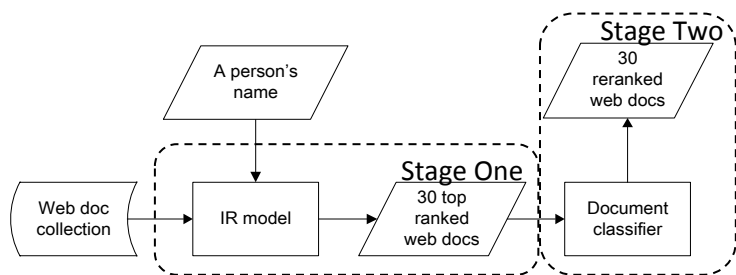


FIGURE 8-11. THE TWO-STAGE ARCHITECTURE

Both of the two architectures above are slightly different from what have been used in retrieving procedures. In the query expansion approach, we used both the specific part (i.e., a person's name) and the generic part (i.e., 'biography') to match retrieve the initial set of documents; in contrast, when retrieving procedures, we only used the specific part (i.e., the content of the goal) to retrieve an initial set which are then filtered by the generic part (i.e., 'how to'). In addition, here we used top 5 documents to generate terms for query expansion; in contrast, when retrieving procedures, we used 10 documents. The reason is because we observed that the top 10 ranked documents contain a lot of irrelevant documents. For the same reason, we choose to use 30 search terms for query expansion instead of 40 in the procedural experiment. In the second stage of the two-stage architecture, the trained classifier first classifies the 30 retrieved documents into two sets, a biographical set and a non-biographical set; then it moves biographical documents up in the list in the way that ensures any biographical document is ranked higher than any non-biographical document while the order between any two biographical documents or any two non-biographical documents remains the same. In contrast, when retrieving procedures, we used the score of the procedural classifier to directly re-rank the retrieved documents.

We used a sample set of the SPIRIT corpus (contained 5778 English web pages) freely available on the Web to construct the web document collection. This collection is small and it is unlikely to contain relevant documents for our biographical questions. We searched Google and downloaded 40 documents for each question, which were then added into the collection. The 40 documents contain 20 top-ranked web pages in Google's retrieval result and 20 web pages randomly selected from the top-ranked 1000. The reason why we used some less high ranked web pages is because they constitute a set of noisy documents in the collection and it is a more challenging task to detect their relevancy than to judge those that are either clearly relevant or completely irrelevant.

8.3.4.3 IR MODELS

For the above-mentioned IR model, we used the PL2 and the BM25 algorithms from the Terrier IR platform⁶⁷. We also used the Bo1 Algorithm provided in the same package to select terms for query expansion.

8.3.4.4 RESULT

We tested 10 systems, which could be organised into two sets. The first set uses BM25 algorithm as the basic IR model and the second set uses PL2 as the basic IR model. Each set includes five systems: a system that only uses the specific part of the questions to query the IR model, a system that uses both the specific part and the generic part of the questions to query the IR model, a system that uses query expansion technique following the architecture in figure 8-10 and two systems that apply the two-stage architecture (one uses the Naive Bayes classifier with featureset1 and another one with featureset3). Each system retrieved an ordered list of 30 documents.

We now need to build the golden-standard retrieval results for each question. We presume documents that are relevant to a question can only be from two sources: the 40 documents that are downloaded from the web according to Google's searching result and all the documents retrieved by different systems. We asked human subjects to go through each document to judge its relevance to the corresponding question.

The mean average precision (MAP)⁶⁸ of different systems are shown in figure 8-12 and table 8-14, where FS1 refers to featureset1 and FS3 refers to featureset3.

From the below table, we can see that in both sets, the first system performs better than the second and the third but worse than the last two systems. We performed pairwise t-tests to test the significance of the differences. The results are shown in table 8-15 and table 8-16. The first number in each cell is the t value and the second number is the p value (level of significance). Numbers that represent significant difference ($p < 0.5$) are in bold face.

⁶⁷ <http://ir.dcs.gla.ac.uk/terrier/index.html>

⁶⁸ The average precision of a single question is the mean of the precision scores after each relevant document is retrieved. The mean average precision is the mean of the average precisions of a collection of questions.

Chapter 8 – Automatically Retrieving Relevant Documents for Procedural and Biographical Questions

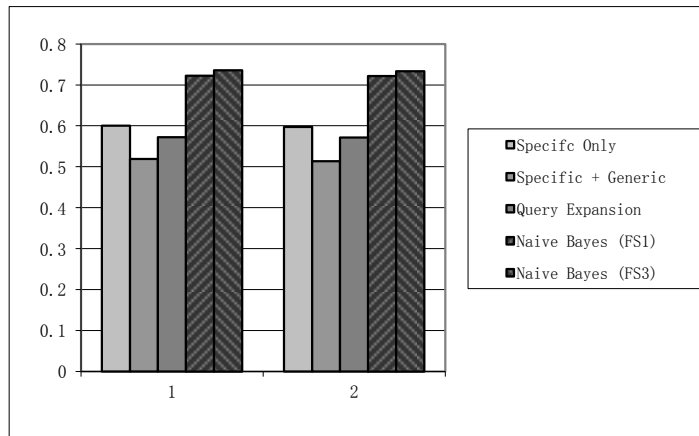


FIGURE 8-12. MAPS OF DIFFERENT SYSTEMS: 1 REFERS TO BM25, 2 REFERS TO PL2

	Systems	MAP
	BM25 + Specific Part	.60066
	BM25 + Specific Part + Generic Part	.51866
BM25	BM25 + Query Expansion	.57195
	BM25 + Naive Bayes (FS1)	.72292
	BM25 + Naive Bayes (FS3)	.73636
	PL2 + Specific Part	.59719
	PL2 + Specific Part + Generic Part	.51321
PL2	PL2 + Query Expansion	.57161
	PL2 + Naive Bayes (FS1)	.72191
	PL2 + Naive Bayes (FS3)	.73379

TABLE 8-14. MAPS OF DIFFERENT SYSTEMS

	BM25+SP	BM25 + SP + GP	BM25+QE	BM25+NB (FS1)	BM25+NB (FS3)
BM25 + SP	-	1.808/.104	.362/.726	-3.104/.013	-3.331/.009
BM25 + SP + GP		-	1.345/.211	-6.533/.000	-6.085/.000
BM25 + QE			-	-2.768/.022	-2.687/.025
BM25 + NB (FS1)				-	-
BM25 + NB (FS3)					-

TABLE 8-15. PAIRWISE T-TESTS RESULTS (BM25)

Chapter 8 – Automatically Retrieving Relevant Documents for Procedural and Biographical Questions

	PL2+SP	PL2 + SP + GP	PL2+QE	PL2+NB (FS1)	PL2+NB (FS3)
PL2 + SP	-	3.225/.010	.825/.431	-3.041/.014	-3.235/.010
PL2 + SP + GP		-	-3.279/.010	-8.140/.000	-7.418/.000
PL2 + QE			-	-4.822/.022	-4.229/.002
PL2 + NB (FS1)				-	-
PL2 + NB (FS3)					-

TABLE 8-16. PAIRWISE T-TESTS RESULTS (PL2)

The results of the pairwise t-tests show that the two-stage architecture (represented by the last two systems in each set) performs significantly better than other three approaches (represented by the first three systems in each set). The difference between approach one (querying an IR model with only the specific part of the questions) and approach two (querying an IR model with both parts of the questions) is significant when using the PL2 model, but not so when using the BM25 model.

8.3.5 DISCUSSION

Compared to retrieving procedures, here we have done comparison among more different approaches. In particular, we have tested whether using the generic part to directly match retrieve documents will bring noise. The experiment results partly support this statement, with significant drop compared to only apply the specific part to match retrieve document when using the PL2 ranking formula.

We also see that the query expansion approach does not outperform the two-stage approach, which is different from what was drawn in section 8.3.4.4. It is probably due to the difference in applying the query expansion (QE) approach as discussed in section 8.3.4.2. If so, it again proves that it is not appropriate to directly use the generic part to match retrieve relevant documents.

8.4 SUMMARY

This chapter experiments on retrieving relevant documents for procedural and biographical questions, both are representative of the fourth question class introduced in section 4.2. We suggest a two-stage approach for this question class, with the first stage focusing on the specific part and the second stage focusing on the generic part using a text classification approach. We show that it performs significantly better than the baseline approach for both procedural questions and biographical questions. The two-stage approach also shows significant gain compared to a query expansion approach in retrieving biographies. The

*Chapter 8 – Automatically Retrieving Relevant
Documents for Procedural and Biographical Questions*

success of the two-stage approach again suggests extended topics do have a significant internal structure.

We also introduce an adapted naïve bayes classification model in this chapter. Compared to basic Bayesian classification model, this model focuses on modelling the difference between two classes instead of modelling the representativeness of a class.

CHAPTER 9

CONCLUSIONS AND FUTURE WORK

9.1 OVERVIEW OF CONTRIBUTIONS

This thesis focuses on studying topic structure and applying the insights drawn from the study to help information retrieval (IR).

The research in the thesis is motivated by the observation that some user queries contain two parts, one of which cannot be effectively retrieved by traditional keyword-matching-based IR approach. Therefore, we aim to improve information retrieval on such kind of queries by studying the query structure. User queries to a document repository resemble a widely discussed notion of topic in the sense that they are both a concise formulation of the discourse content. Observations on topic expressions also indicate that they contain two different parts with different roles. Therefore, instead of only studying user queries, we broaden the research to focus on topics. The contributions of the thesis could be divided into two aspects: a theoretical aspect and a practical aspect.

On the theoretical side, the main contribution is to the theory of topic. Specifically, we reviewed existing theories of topic and developed a topic classification schema (section 9.1.1). We conducted a series of experiments on topic expressions and developed the theory of extended topic to better characterise the structure of topic (section 9.1.2). Specifically, we concluded that an extended topic contains two parts, a generic part and a specific part. We also advanced the understanding of how the different parts in extended topic are derived from the elements in a relevant discourse (section 9.1.3). Many terms could appear both in the generic part and in the specific part of a topic expression, but when the terms appear in different positions, people would have different expectations of the discourse content. We explained this phenomenon by suggesting that there are *two modes of being a topic* (9.1.4).

On the practical side, the main contribution is to the research of information retrieval. Specifically, we listed the problems when applying the current IR techniques to retrieve the generic part of extended topic. We proposed different improvements to existing IR techniques to better tackle different types of generic topics (section 9.1.5). The focus on analysing topic structure forms a new research angle in the research area of information retrieval.

Below in section 9.1.1 to 9.1.5, we will talk in more details about the above-mentioned contributions. Section 9.1.6 summarises a few more minor contributions that we have made.

9.1.1 A TOPIC CLASSIFICATION SCHEMA AND THE NATURE OF INDICATIVENESS

The notion of topic is widely discussed in theoretical linguistics and in applied natural language processing research. In chapter 2, we identified three views of this notion from the literature, including an important sentential element in sentence structure analysis, a concise representation of discourse content and a discourse organisation principle. The second view is the one we are concerned about here. We further distinguished between two types of topic: *indicative topic* and *informative topic*. Below is the formal definition of the distinction between *indicativeness* and *informativeness*, with special emphasis on defining the nature of *indicativeness*.

One view of the topic of a discourse is that it is a concise formulation of the discourse content. Under this view, there are two types of topic, indicative topic and informative topic. An informative topic aims to reproduce the core information in the discourse. In contrast, an indicative topic only indicates the content in the discourse without imparting any detailed information. Specifically, an indicative topic defines a range of information that the discourse contains, just like a natural language query to document repositories, which defines the range of information that is relevant. We assume here that information in a discourse is represented in a formal knowledge representation framework in which knowledge comprises a network of entities linking with each other by relationships. In term of this model, the nature of indicativeness is defined as a function which, given a knowledge network, primes a relevant subnetwork.

9.1.2 STRUCTURE OF EXTENDED TOPIC

In Chapter 3, we analysed how indicative topic fulfils the function of selecting a relevant subnetwork from a knowledge base. We identified two different parts in an indicative topic which play different roles in fulfilling this function. The theory of extended topic is developed to formalise the structure of indicative topic.

An extended topic is a concise, indicative formulation of the content of a discourse. Still using the above-mentioned knowledge representation model in which knowledge comprises a

Chapter 9 – Conclusions and Future Work

network of entities and relationships, an extended topic identifies a relevant subnetwork by the following strategy: first, pick an entity (or entities); second, apply a rule for navigating from this entity. The components of an extended topic are accordingly the focused entity (which identifies the entity from which we navigate) and the perspective (which identifies the navigation rule).

There are different types of perspectives. The simplest type of perspective is a single relation between the focused entity and an unspecified entity, such as ‘eat’ in topic ‘what does cow eat?’. Slightly more complex perspectives include a single relationship and a general category which denotes the kind of unspecified entities. For example, in topic ‘animals that eat grass’, ‘eat’ is a relationship and ‘animals’ represents a general category. Even more complex perspectives would define a category of relationships/facts. For example, in topic ‘procedure for cooking rice’, ‘procedure’ defines a list of steps; in topic ‘the anatomy of the jaw’, ‘anatomy’ defines compositional and positional relationships. We refer to the general category in the latter two types of perspectives as the generic part and the particular relationship in the first two types of perspectives together with the focused entity as the specific part. It is the generic part that is insufficiently addressed in previous work on topic.

Chapter 3 also provides a set of contexts where extended topic is explicitly formulated, including WH-questions, sentences describing the plan of a discourse, etc. We delivered a set of empirical studies on topic expressions in chapter 4 to verify the above-mentioned distinction between the specific topic and the generic topic.

In section 4.2, we experimented with a small set of WH-questions in the medical domain. We showed that:

WH-questions could be transformed into the form of ‘what is/are the <G> of/that/for <S>’, where G corresponds to the generic part and S corresponds to the specific part.

In section 4.3, we verified whether the concepts collected from the generic part of WH-questions could work as a short-hand for a family of facts/relationships. The experiment result shows that given a general concept and two passages, with one passage being relevant to the concept and one irrelevant, people could reliably pick up the relevant one. Here neither of the two passages explicitly contains the general concept. This result shows that:

given a general concept, people do have a shared view on whether a passage is relevant or not. In other words, general concepts could function in real life as knowledge selecting criteria.

In section 4.4, we acquired a set of topic expressions from academic papers. These topic expressions take the form of 'this paper describes/presents the/a <G> of/that/for <S>', where the G part corresponds to the generic part and the S part corresponds to the specific part. We extracted the nouns in the G part and in the S part and put them into two groups. We showed that:

The generic part of topic expressions contains more general terms than the specific part.

All of the above findings comply with the theory of extended topic.

9.1.3 RELATIONSHIP BETWEEN EXTENDED TOPIC COMPONENTS AND DISCOURSE CONSTITUENCIES

Above we noted that extended topic contains two parts, a generic part and a specific part. The generic part is a general category that specifies the kind of detailed information related to the specific part. Thus, in a discourse relevant to an extended topic, the specific part will remain and the generic part will be replaced by detailed information belonging to the category that it denotes. For a user query in the form of extended topic, a relevant discourse would contain the specific part (of the query) which is *known* to the questioner and the detailed information that the generic part (of the query) denotes which is *unknown* to him/her. The *known/unknown* distinction is also known as the *given/new* distinction defined in the information structure theory. While the information structure theory is typically applied at the sentence level, here we extend the theory to apply to the discourse level; therefore, the information in a discourse could be divided into two parts: the *given* part and the *new* part. In section 4.5, we presented experiments to verify the relationship between the generic/specific part of an extended topic and the given/new part in a relevant discourse. We divided topic expressions into generic parts and specific parts and divided discourses into given parts and new parts. For each concept in the topic expressions, we generated a list of keywords associated with the concept as its *signatures*. We show that the *signatures* of generic topic appear more in the *given* part of a discourse than the *signatures* of specific topic. From the experiment result, we could derive the relationship between different parts of extended topic and different elements in a relevant discourse, as below:

In a discourse that contain relevant information to an extended topic, the specific part of the topic corresponds to the given information in the discourse and the generic part is a general characteristic of the new information in the discourse.

9.1.4 TWO MODES OF BEING A TOPIC

As mentioned in section 3.2 and section 5.2, many concepts could work both as a generic topic and as a specific topic in different topic expressions. For example, the concept ‘animal’ is a specific topic in topic expression ‘the origin of animals’ and is a generic topic in topic expression ‘the animals that eat grass’. People would have different expectations of the discourse content for a concept being a generic topic as opposed to being a specific topic. For example, for ‘animals that eat grass’, people would expect to have detailed description about the kind of animals that eat grass; in contrast, in a discourse about the ‘origin of animals’, people would expect to see the origin of animals as a whole, only listing concrete animal names if different animals have different origins. Furthermore, we may take the whole topic expression ‘animals that eat grass’ as a specific topic; in this case, we would expect the discourse to talk about some common attributes about this type of animals (i.e., herbivorous animal). Thus, we conclude that

There are two modes of being a topic: (being talked about) as generic topic and (being talked about) as specific topic. In other words, generic topic and specific topic are two modes of being a topic. For a particular concept, when being a generic topic of a discourse, it would work as a semantic category and the discourse would contain detailed knowledge that falls into this category; when being a specific topic of a discourse, it would be treated as referring to one concept and the discourse would talk about knowledge that relates to this concept as a whole.

9.1.5 APPLYING EXTENDED TOPIC TO INFORMATION RETRIEVAL

According to the conclusion made in the last section, the generic part of extended topic will be replaced by the new information in a relevant discourse; thus, typical key-word-matching-based IR techniques are not suited to retrieve the generic part. In section 5.6, we propose a general method to sharpen the retrieval of the generic part. Specifically, since many general concepts either retrieve an entity/event/proposition by its relationship with other entity/event/proposition or prime a family of relationships, so we propose to improve detecting relationships using discourse connectives, cue phrases and prepositions. There are some variations in the detailed design to tackle different types of general concepts. We experimented on retrieving a few different types of general concepts in chapter 7 and chapter 8.

In chapter 7, we experimented on retrieving relevant documents for causal questions. Causal questions represent the type of extended topic that retrieves an entity/proposition/event that has a particular relationship (in this case, 'cause') with a specified entity/preposition/event. We compiled 11 causal indicators (e.g., cue phrases or prepositions) to detect the causal relationship. We also invented a weighting schema on basis of the vector space model to emphasize that the causal indicators must occur within a certain distance to the specific topic. This method is compared against typical key-word-matching-based IR techniques⁶⁹ (the default vector space model and the BM25 model implemented in Lucene-1.4.2 package) and performs statistically significantly better than the baseline systems.

In chapter 8, we experimented on retrieving relevant documents for procedural and biographical questions, both of which represent the type of extended topic that retrieves a family of facts/relationships. We proposed a two-stage IR approach to tackle this type of extended topic. At the first stage, we used the specific topic to retrieve a list of top ranked documents; at the second stage, we trained a text classifier to re-rank the top ranked documents based on the generic topic. The two-stage approach is again compared against the typical key-word-matching-based IR approach⁷⁰. Experiment results show that the proposed two-stage approach performs statistically significantly better than the baseline systems.

The conclusion below is drawn from the experiment results.

The proposed general method, which uses discourse connectives, cue phrases and prepositions to improve retrieving the generic part of extended topic, performs statistically significantly better than the key-word-matching-based IR techniques.

9.1.6 OTHER CONTRIBUTIONS

In addition to the contributions listed above, the thesis also increases knowledge or brings technical advances in a few related theoretical and application areas, summarised as below:

- From theoretical point of view, the notion of extended topic also improves the knowledge of how a discourse is organised. Specifically, it is a common structure of a discourse to

⁶⁹ the default vector space model and the BM25 model implemented in Lucene-1.4.2 package

⁷⁰ Same as 1

contain several passages with a shared specific topic, each focusing on a different *perspective* of the specific topic.

- From practical point of view, in addition to information retrieval, we also propose ways in chapter 5 to apply extended topic to help other application areas, including knowledge organisation and text indexing, text segmentation, discourse planning and generating indicative topic expressions.
- In the experiment on classifying procedural texts, we adapted Naive Bayes model to focus more on modelling the difference between two classes. Experiment results show that the adapted Naive Bayes model performs better than Naive Bayes classifier when combined with very distinctive features.

9.2 FUTURE WORK

The main contributions of this thesis include developing a theory of extended topic to better characterise the structure of user queries and proposing ways this theory can be used to improve information retrieval. Below we will present a few future research directions along this line in section 9.2.1 to 9.2.3. In addition to information retrieval, the theory of extended topic could also be applied to other application areas, which will be presented in more details in section 9.2.4.

9.2.1 MODEL OTHER GENERIC CONCEPTS TO IMPROVE INFORMATION RETRIEVAL ON THEM

One main insight from the theory of extended topic is that the generic part of extended topic will be replaced by the new information in the discourse. Thus, typical match retrieval approach could not effectively retrieve documents relevant to the generic part because they will not appear in the discourse. Since many generic topics denote a family of relationships, we proposed to detect generic topics by identifying the relationships using shallow linguistic or lexical features, such as discourse connectives, prepositions and cue phrases. In chapter 8, we trained classifiers to identify biographies and procedures using shallow features; the classifiers are then combined into a two-stage retrieval framework to improve retrieval for biographical and procedural questions. As a future work, we propose to model some other generic topics using the same approach.

The problem with this approach is that a classifier must be built for each generic topic and there is no generally applied model. However, many generic topics represent topic areas of

broad interests. For example, in medical domain, it is very common for people to ask about the *prevention* of a certain disease and the *side effect* of a certain medicine. In question answering, a research area closely related to information retrieval, there are studies specially targeting at answering one particular type of questions, such as questions asking about *opinions* and *definitions*. Therefore it is worthy of the effort to develop a classifier for each generic topic.

9.2.2 STUDY THE DIFFERENCE IN DOCUMENT DISTRIBUTION BETWEEN GENERIC TOPIC AND SPECIFIC TOPIC

Many IR models weight terms based on term document distribution, such as the idf term weighting schema and the 2-Poisson term weighting schema. As noted in section 6.2.8, since the generic topic of an extended topic is likely to be hidden in relevant documents, it may be over-weighted using the typical term weighting schema. As a future research direction, one could further investigate the difference between the document distribution of generic topic and specific topic. Below are some concrete ideas.

Many IR models weight a term by evaluating how it deviates from a random distribution in a document collection. The above-mentioned idf term weighting schema and the 2-Poisson model are both models of randomness. However, in section 4.3, we show that generic topics contain more scientific research general terms while specific topics contain more scientific research specific terms. So we propose to study whether specific topics follow a random distribution only within a certain scientific research domain and generic topics follow random distribution across all the domains. If the answer to the above hypothesis is yes, then we propose to improve information retrieval by first classifying documents into different domains and then applying the term weighting schema only based on documents in the domain that the question falls into. Instead of using scientific research domains to classify documents, one could also investigate on other document classification criteria.

The above approach studies the difference between terms typically used in generic topics and those typically used in specific topics in terms of document distribution. However, according to the theory of extended topic, generic topic and specific topic are two different roles in the function of selecting relevant information; a concept could appear both as a generic topic and as a specific topic in different extended topics. For example, the concept 'mammal' is part of the generic topic in 'the mammal that eats bamboo'; in contrast, it is part of the specific topic in 'the eating habit of mammals'. The 2-Poisson term weighting schema presumes that a

document collection could be divided into two sets based on a content-bearing term: an elite set⁷¹ and a non-elite set. The elite set is relevant to the term. For a term in the query, it models the term distribution in the elite set and the non-elite set separately. To score a document, it calculates the probability for the document to belong to the elite set based on the term frequencies. Based on the theory of extended topic, the elite set could be further divided into two subsets: *documents that are relevant to a specific topic ‘mammal’* and *documents that are relevant to a generic topic ‘mammal’*. Then for the query ‘the mammal that eats bamboo’, the document scoring method should be revised to be based on the probability for the document to belong to the set that are relevant to a generic topic ‘mammal’ and the probability for the document to belong to the set that are relevant to a specific topic ‘eats bamboo’. Many IR systems approach an estimate of the elite set for a term by simply taking all the documents that contain the term. It would be hard to further separate the elite set into two subsets. One solution is to detect whether the term is presented as given information or new information in a discourse. Specifically, we can extract all the documents that contain term ‘mammal’ as given information. These documents are taken as relevant to ‘mammal’ as a specific topic; the rest of the documents in the elite set are taken as relevant to ‘mammal’ as a generic topic.

In the solution proposed above, we apply the mapping relationship between the generic/specific part of an extended topic and the given/new part in a relevant discourse. Below in section 9.2.3 we will further talk about this direction.

9.2.3 RELATE DIFFERENT PARTS OF EXTENDED TOPIC TO DIFFERENT DISCOURSE CONSTITUENCIES

One of the findings of the thesis is that the generic part of an information request will be replaced in the answer by information that are *new* to the questioner, while the specific part will be kept as a part that are *known*. Therefore, to match a document to an information request, we could use the given information (i.e. the known information) in the discourse to match the specific part of the information request and use the new information to match the generic part. This idea that relates different parts in an information request to different

⁷¹ The concept of *eliteness* is proposed by Harter (1975a), Robertson and Walker (1997) and Robertson et al. (1993, 1996). According to Robertson et al. (1993, 1996), the “elite” set for a particular term may be interpreted to mean those documents which can be said to be “about” the concept represented by the term. According to Robertson and Walker(1997), those documents which are “about” this concept are described as “elite” for the term.

discourse constituencies has never been applied in the research of information retrieval. One example presented in section 5.6 is to apply the mapping relationship to improve the language model approach.

There are several problems in this research direction. First, as illustrated in section 4.5, it is hard to divide the given information and the new information in a discourse using simple linguistic features. Secondly, the above mapping relationship assumes that the new information to the questioner is also what the author meant to be new, which is not always true. For example, in a discourse about Pandas, the sentence 'Pandas eat bamboo' contains a given entity 'Pandas' and a 'eat bamboo' which is meant to be new information. The sentence could answer question 'what animal eats bamboo?'. We see that in this question, 'animal' is the generic part which points to the unknown information. So to the questioner, 'Pandas' is the new information. The above problem is most likely to happen for factoid questions (i.e., questions that seek for a single piece of fact) but less so for questions that ask about a lot of detailed information which could only be delivered in a long discourse. For example, for the question 'what are the consequences of modernisation', it is unlikely to have a discourse that includes all the detailed consequences of modernisation as given information. Thirdly, it is hard to model the relationship between the generic topic and the new information in a relevant discourse since it is not the generic topic itself that will occur as new information in the discourse. On the contrary, similar to the specific topic, the concept in the generic topic is likely to occur in the given part in a relevant discourse. For example, in a discourse about 'the weather of England', we may have a sentence 'The weather of England is not nice', in which 'weather' is a given concept. In spite of these problems, the above proposed direction is still worth trying.

9.2.4 OTHER FUTURE WORK

Above we focus on the future work in applying the theory of extended topic to improve information retrieval (IR). In addition to IR, the theory of extended topic could also be applied to many other applied research areas. For example, extended topic would define good knowledge organisation schemata, with one part representing a central entity and another part representing one perspective of the entity. Such a structure resembles the structure of user queries and therefore could be directly matched to user queries to find relevant pieces of knowledge. In chapter 5, we also discussed in details about how to apply extended topic to discourse segmentation, discourse planning and generating indicative topic expressions. These

Chapter 9 – Conclusions and Future Work

research directions are also worth pursuing in future. From the theoretical side, the theory of two modes of being a topic has not been fully explored. As a future research direction, one could design survey to test whether human subjects would have different expectations of the discourse content for a given topic in different modes.

REFERENCES

- Ades, A. and M. J. Steedman (1982). On the order of words. *Linguistics and Philosophy* 4, pp. 517–558.
- Alhadi, A., T. Gottron, J. Kunegis and N. Naveed (2011). LiveTweet: Microblog Retrieval Based on Interestingness and an Adaptation of the Vector Space Model. In *Proceedings of 20th Meeting of the Text Retrieval Conference (TREC 2011)*.
- Amati, G. and C. J. van Rijsbergen (2002). Probabilistic models of information retrieval based on measuring the divergence from randomness. *ACM Transactions on Information Systems*, 20 (4), pp. 357-389.
- Ammann, H. (1928). *Die menschliche Rede 2. Der Satz*. Darmstadt, Germany.
- Appel, R., H. Komorowski, C. Barr and R. Greenes (1988). Intelligent focusing in knowledge indexing and retrieval – the relatedness tool. In *Proceedings of SCAMC'88*, pp. 152 – 157. IEEE, 1988.
- Appelt, D. and D. Israel (1999). Introduction to Information Extraction Technology. In *Proceedings of International Joint Conference on Artificial Intelligence (IJCAI-99)*.
- Arora, R. and B. Ravindran (2008). Latent Dirichlet Allocation and Singular Value Decomposition Based Multi-document Summarization. In *Proceedings of the 8th IEEE International Conference on Data Mining (ICDM 2008)*.
- Baeza-Yates, R. A. and B. Ribeiro-Neto (1999). *Modern Information Retrieval*. New York: ACM Press, 1st edition.
- Baeza-Yates, R. A. and B. Ribeiro-Neto (2011). *Modern Information Retrieval*. Addison Wesley, 2nd edition, 2011.
- Barzilay, R. and M. Elhadad (1997). Using lexical chains for text summarization. In *Proceedings of the Workshop on Intelligent Scalable Text Summarization at the ACL/EACL Conference*, pp. 10–17. Madrid, Spain.
- Berger, A. and J. Lafferty (1999). Information retrieval as statistical translation. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development of Information Retrieval (SIGIR-99)*, pp. 222-229.
- Blair-Goldensohn, S., K. R. McKeown and A. H. Schlaikjer (2003). *A Hybrid Approach for Answering Definitional Questions*. Technical Report CUCS-006-03, Columbia University, 2003.

References

- Blei, D., A. Ng and M Jordan (2003). Latent dirichlet allocation. In *the Journal of Machine Learning Research*, vol. 3, pp. 993-1022.
- Bobrow, D., R. Kaplan, M. Kay, D. Norman, H. Thompson and T. Winograd (1977). GUS, a frame driven dialog system. *Artificial Intelligence*, 8, pp. 155-173.
- Boguraev, B. and C. Kennedy (1997). Saliency-based content characterisation of text documents. In *Proceedings of the ACL/EACL Workshop on Intelligent Scalable Text Summarization*.
- Bookstein, A. and D. R. Swanson (1974). Probabilistic models for automatic indexing. *Journal of the American Society for Information Science* 25 (5), pp. 313–318.
- Bozsahin, H. Cem and N. V. Findler (1992). Memory-based hypothesis formation: Heuristic learning of commonsense causal relations from text. *Cognitive Science*, 16(4), 431-454
- Brill, E. (1993). *A Corpus-Based Approach to Language Learning*. Ph.D. thesis, University of Pennsylvania, Philadelphia, PA, 1993.
- Brill, E. (1994). Some advances in transformation-based part of speech tagging. In *Proceedings of the Twelfth national Conference on Artificial Intelligence*, pp. 722-727.
- Brown, G. and G. Yule (1983). *Discourse Analysis*. Cambridge University Press.
- Burges, C. et al. (2005). Learning to rank using gradient descent. In *Proceedings of ICML '05*, pp. 89–96.
- Burges, C., R. Ragno and Q. Le. (2006). Learning to rank with nonsmooth cost functions. In *NIPS*, pp. 193.
- Burke, R. D., K. J. Hammond, V. A. Kulyukin, S. L. Lytinen, N. Tomuro and S. Schoenberg (1997). Question answering from frequently-asked question files: experiences with the FAQ finder system. *AI Magazine*, 18(2), pp. 57-66.
- Carpenter, B. (2004). Phrasal queries with LingPipe and Lucene. In *Proceedings of the 13th Meeting of the Text Retrieval Conference (TREC)*. Gaithersburg, Maryland.
- Chang, Y.-L. and J.-T. Chien (2009). Latent dirichlet learning for document summarization. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2009)*.

References

- Chiaramella, Y. and J. Nie (1990). A retrieval model based on an extended model logic and its application to the RIME experimental approach. In *Proceedings of the 13th International ACM SIGIR Conference on Research and Development on Information Retrieval*, pp. 25-43.
- Clarke, C., G. Cormack, M. Laszlo, T. Lynam and E. Terra (1998). The impact of corpus size on question answering performance. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in IR*. Tampere, Finland.
- Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20, pp. 37-46.
- Cohen, D., E. Amitay and D. Carmel (2007). Lucene and Juru at Trec 2007: 1-Million Queries Track. In *Proceedings of the 16th Text Retrieval Conference (Trec 2007)*.
- Croft, B., D. Metzler and T. Strohman (2009). *Search Engines: Information Retrieval in Practice*. Pearson, 2009.
- Croft, W. B., H. R. Turtle and D. D. Lewis (1991). The use of phrases and structured queries in information retrieval. In *Proceedings of the 14th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 32-45.
- Daneš, F. (1968). Some thoughts on the semantic structure of the sentence. *Lingua*, 21, pp. 55-69.
- Dang, H.-T., D. Kelly and J. Lin (2007). Overview of the TREC 2007 question answering track. In *Proceedings of the Sixteenth Text Retrieval Conference (TREC 2007)*.
- Dang, H.-T., J. Lin and D. Kelly (2006). Overview of the TREC 2006 question answering track. In *Proceedings of the Fifteenth Text Retrieval Conference (TREC 2006)*.
- de Beaugrande, R. (1980). *Text, Discourse and Process*. London: Longman.
- Davies, N. (1999). *The Isles: A History*. New York: Oxford University Press, 1999.
- Dejong, G. (1982). An overview of the FRUMP system. In: Lehnert, W. G. and Ringle, M. H. (eds.) *Strategies for Natural Language Processing*, pp. 149-176. Hillsdale, New Jersey: Erlbaum.
- Delin, J., A. Hartley, C. Paris, D. Scott and K. V. Linden (1994). Expressing Procedural Relationships in Multilingual Instructions. In *Proceedings of the 7th International Workshop on Natural Language Generation*.

References

- Duchastel, J., L. –C. Paquin and J. Beauchemin (1992). Automated syntactic text description enhancement: The thematic structure of discourse utterances. *Computers and the humanities*, issue 1, 1992.
- Elworthy, D. (2000). Question answering using a large NLP system. In *Proceedings of the Ninth Text Retrieval Conference (TREC-9)*, pp. 355-360.
- Fagan, J. L. (1987). Automatic phrase indexing for document retrieval: An examination of syntactic and non-syntactic methods. In *Proc. 10th Ann. Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval*, pp. 91-101.
- Ferrucci, D. A. (2012). Introduction to “This is Watson”. In *IBM Journal of Research and Development*, vol. 56, no. 3/4, pp. 1-15.
- Ferrucci, D. et al. (2010). Building Watson: An overview of the deep QA project. In *AI Magazine*, vol. 31, no. 3.
- Firbas, J. (1964). On defining the theme in functional sentence perspective. *Travaux Linguistique de Prague*, pp. 267-280.
- Firbas, J. (1971). On the concept of communicative dynamism in theory of Functional Sentence Perspective. *Brno Studies in English*, 7, pp. 12-47.
- Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76, pp. 378–382.
- Freund, Y. and L. Mason (1999). The alternating decision tree learning algorithm. In *Proceedings of the Sixteenth International Conference on Machine Learning*, pp. 124-133. Bled, Slovenia.
- Fuhr, N. (1992). Probabilistic models in information retrieval. *The Computer Journal*, 35(3), pp. 243-255.
- Girju, R. (2003). Automatic detection of causal relations for question answering. In the *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL 2003)*, Workshop on "Multilingual Summarization and Question Answering - Machine Learning and Beyond".
- Givon, T. (1983). Topic continuity in spoken discourse: A quantitative cross-language study. *Typological Studies in Language*, Volume 3. Amsterdam: John Benjamins.

References

- Givon, T. (1988). The pragmatics of word order: predictability, importance and attention. In: Hammond, M., Edith Moravcsik, and Jessica Werth (eds.) *Studies in Syntactic Typology*, pp. 243-284. Amsterdam: John Benjamins.
- Gong, Y. and X. Liu (2001). Generic Text Summarization Using Relevance Measure and Latent Semantic Analysis. In *proceedings of the 24th annual international ACM SIGIR conference on research and development in information retrieval*.
- Gonzalo, J., F. Verdejo, I. Chugur and J. Cigarran (1998). Indexing with WordNet synsets can improve Text Retrieval. In *Proceedings of the COLING/ACL '98 Workshop on Usage of WordNet for NLP*. Montreal.
- Green, B. F., A. K. Wolf, C. Chomsky and K. Laughery (1961). BASEBALL: An automatic question answerer. In *Proceedings of the Western Joint Computer Conference*, 19, pp. 219-224.
- Greengrass, E. (2001). *Information retrieval: A survey*. DOD Technical Report TR-R52-008-001.
- Grosz, B. J., A. K. Joshi and S. Weinstein (1995). Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, 12(2), pp. 203-225.
- Grosz, B. J. and C. L. Sidner (1986). Attention, intentions and the structure of discourse. *Computational Linguistics*, 12, pp. 175-204.
- Gruber, T.R. (1993). A translation approach to portable ontology specifications. *Knowledge Acquisition*, 5(2), pp. 199–220.
- Gundel, J. K. (1974). *The role of topic and comment in linguistic theory*. University of Texas, Ph.D. dissertation.
- Gundel, J. K. (1988). Universals of Topic-Comment Structure. In: Hammond, M., Edith Moravcsik, and Jessica Werth (eds.) *Studies in Syntactic Typology*, pp. 209-239. Amsterdam: John Benjamins.
- Gundel, J. K. (2003). Information structure and referential givenness/newness. How much belongs in the grammar? *Journal of Cognitive Science*, 4, pp. 177-199.
- Gundel, J. K., N. Hedberg and R. Zacharski (1993). Cognitive status and the form of referring expressions in discourse. *Language*, 69, pp. 274-307.
- Hackl, R., R. Kölle, T. Mandl, A. Ploedt, J. -H. Scheufen and C. Womser-Hacker (2004). Multilingual Retrieval Experiments with MIMOR at the University of Hildesheim. In C. Peters, M. Braschler, J. Gonzalo, M. Kluck (eds.), *Comparative Evaluation of Multilingual Information*

References

- Access Systems: 4th Workshop of the Cross-Language Evaluation Forum, CLEF 2003, Revised Selected Papers*, pp. 166-173. Springer [LNCS 3237], Trondheim, Norway.
- Hahn, U. (1990). Topic parsing: accounting for text macro structures in full-text analysis. *Information Processing and Management*, 26, pp. 135-170.
- Halliday, M. (1967). Notes on Transitivity and Theme in English II. *Journal of Linguistics*, 3, pp. 199-244.
- Halliday, M. and R. Hasan (1976). *Cohesion in English*. London, Longman, 1976.
- Harabagiu, S., D. Moldovan, P. Marius, S. Mihai, M. Rada, G. Roxana, R. Vasile, L. Finley, M. Paul and B. Razvan (2001). Answering Complex, List and Context Questions with LCC's Question-Answering Server. *Tenth Text REtrieval Conference (TREC-10)*. Gaithersburg, MD.
- Harter, S. P. (1974). *A Probabilistic Approach to Automatic Keyword Indexing*. Ph.D. Thesis, Graduate Library, The University of Chicago, Thesis No. T25146.
- Harter, S. P. (1975a). A probabilistic approach to automatic keyword indexing. Part I: On the distribution of specialty words in a technical literature. *J. ASIS* 26, pp. 197–216.
- Harter, S. P. (1975b). A probabilistic approach to automatic keyword indexing. Part II: An algorithm for probabilistic indexing. *J. ASIS* 26, pp. 280–289.
- Hearst, M. (1997). TextTiling: Segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, 23 (1), pp. 33-64.
- Heflin, J. and J. Hendler (2000). Searching the Web with SHOE. In *AAAI-2000 Workshop on AI for Web Search*.
- Hiemstra, D. (2001). *Using Language Models for Information Retrieval*. Ph.D. Thesis, Centre for Telematics and Information Technology, University of Twente, January 2001. ISSN 1381-3617 (no. 01-32), ISBN 90-75296-05-3.
- Hirschman, L. and R. Gaizauskas (2001). Natural Language Question Answering: the View from Here. *Journal of Natural Language Engineering, Special Issue on Question Answering*.
- Hobbs, J. R. (1985). *On the Coherence and Structure of Discourse*. CSLI Stanford. Report nr. 85-37.
- Hockett, C. F. (1958). *A Course in Modern Linguistics*. New York: Macmillan.

References

- Hovy, E., L. Gerber, U. Hermjakob, M. Junk and C. -Y. Lin (2001a). Question Answering in Webclopedia. In *Proceedings of the Ninth Text Retrieval Conference (TREC-9)*, pp. 655-664.
- Hovy, E. H., L. Gerber, U. Hermjakob, C. -Y. Lin and D. Ravichandran (2001b). Toward Semantics-Based Answer Pinpointing. In *Proceedings of the DARPA Human Language Technology Conference (HLT)*. San Diego, CA.
- Hovy, E. H. and C. -Y. Lin (1998). Automated Text Summarization in SUMMARIST. In M. Maybury and I. Mani (eds.), *Intelligent Scalable Summarization Text Summarization*.
- Hutchins, J. (1977a). On the problem of 'aboutness' in document analysis. *Journal of Informatics* 1(1), pp. 17-35.
- Hutchins, J. (1977b). On the structure of scientific texts. UEA Papers in *Linguistics* 5, pp. 18-39.
- Ittycheriah, A. and S. Roukos (2002). IBM's Statistical Question Answering System - TREC-11. In *Proceedings of the 11th Text REtrieval Conference*.
- Jin, R., A. G. Hauptmann and C. Zhai (2002). Title Language Model for Information Retrieval. *2002 ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'02)*.
- John, G. and P. Langley (1995). Estimating Continuous Distributions in Bayesian Classifiers. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, pp. 338-345. Morgan Kaufmann, San Mateo.
- Justeson, J. S. and S. M. Katz (1995). Technical Terminology: some linguistic properties and an algorithm for identification in text. *Natural Language Engineering*, 1, 1, pp. 9-27.
- Kan, M., J. Klavans and K. McKeown (1998). Linear Segmentation and Segment Significance. In *Proceedings of 6th International Workshop of Very Large Corpora (WVLC-6)*, pp. 197-205. Montreal, Quebec, Canada.
- Kan, M., K. McKeown and J. Klavans (2001). Domain-specific informative and indicative summarization for information retrieval. In *Proceedings of the Document Understanding Workshop*. New Orleans, USA.
- Katz, B. (1997). Annotating the World Wide Web using natural language. In *Proceedings of the 5th RIAO Conference on Computer Assisted Information Searching on the Internet (RIAO '97)*.
- Katz, B., G. Borchardt and S. Felshin (2006). Natural language annotations for question answering. In *Proceedings of the 19th International FLAIRS Conference (FLAIRS 2006)*.

References

- Katz, B. et al. (2005). External knowledge sources for question answering. In *Proceedings of the 14th Annual Text Retrieval Conference (TREC 2005)*.
- Katz, B. et al. (2007). CSAIL at TREC 2007 question answering. In *Proceedings of the Sixteenth Text Retrieval Conference (TREC 2007)*.
- Katz, J. (1980). Chomsky on meaning. *Language* 56, pp. 1-42.
- Keenan, E. O. and B. Schieffelin (1976). Topic as a discourse notion. In C. N. Li (ed.), *Subject and Topic*, pp. 337-384. New York: Academic Press.
- Kelly, D., V. Murdock, X. J. Yuan, W. B. Croft and N. J. Belkin (2002). Features of Documents Relevant to Task- and Fact-Oriented Questions. In *Proceedings of the Eleventh International Conference on Information and Knowledge Management (CIKM '02)*, pp. 645-647. McLean, VA.
- Kendall, M. (1979). *The Advanced Theory of Statistics*. Fourth Edition. Griffin, London.
- Khoo, C. S. G., S. Chan and Y. Niu (2000). Extracting Causal Knowledge from a Medical Database Using Graphical Patterns. *Proceedings of the ACL-2000*, pp. 336-344.
- Kim, S., H. Seo and H. Rim (2004). Information Retrieval Using Word Senses: Root Sense Tagging Approach, in *Proceedings of the 27th ACM SIGIR*, pp. 258-265. Sheffield, UK.
- Kim, Y.-H. and B.-T. Zhang (2001). Document Indexing Using Independent Topic Extraction. In *Proceedings of the Third International Conference on Independent Component Analysis and Signal Separation*, pp. 557-562.
- Komagata, N. (1999). *A Computational Analysis of Information Structure Using Parallel Expository Texts in English and Japanese*. Ph.D. dissertation. University of Pennsylvania.
- Komagata, N. (2001). Entangled Information Structure: Analysis of Complex Sentence Structures. In *Proceedings of ESSLLI 2001 Workshop on Information Structure, Discourse Structure and Discourse Semantics*, pp.53-66. Helsinki.
- Kruijff-Korbayová, I. and B. L. Webber (2001). Concession, implicature and alternative sets. In *Proceedings of the International Workshop on Computational Semantics (IWCS-4)*. Tilburg.
- Kupiec, J., J. Pedersen and F. Chen (1995). A Trainable Document Summarizer. In *Proceedings of the Eighteenth Annual International ACM Conference on Research and Development in Information Retrieval (SIGIR)*, pp 68-73. Seattle, WA.
- Lafferty, J. and C. Zhai (2001). Document language models, query models, and risk minimization. In *Proceedings of the 24th ACM Conference on Research*, pp. 111-119.

References

- Lehnert, W. (1978). *The Process of Question Answering*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Leidner, J. L., B. Johan, D. Tiphaine, R. C. James, C. Stephen, C. J. Bannard, M. Steedman and B. Webber (2004). The QED open-domain answer retrieval system for TREC 2003. In *Proceedings of the Twelfth Text Retrieval Conference (TREC 2003)*: 595–599. Gaithersburg, MD.
- LeThanh, H., G. Abeysinghe and C. Huyck (2003). Using Cohesive Devices to recognize Rhetorical Relations in Text. In *Proceedings of 4th Computational Linguistics UK Research Colloquium (CLUK-4)*, pp. 123-128.
- Lewis, D. (1997). *Reuters-21578 Text Categorization Test Collection*, Distribution 1.0, README file (Version 1.2), Manuscript, September 26, 1997. Accessed at <<http://www.daviddlewis.com/resources/testcollection/reuters21578/readme.html>>
- Lewis, D. and K. Sparck Jones (1996). Natural Language Processing for Information Retrieval. *Communications of the ACM*, 1996, Vol. 39, N° 1, pp. 92-101.
- Lioma, C. and I. Ounis (2006). Examining the content load of part of speech blocks for information retrieval. In *Proceedings of the COLING-ACL '06*.
- Lin, C.-Y. and E. Hovy (1998). *Automated Text Summarization and the SUMMARIST System*. TIPSTER III Final Report (SUMMAC). ISI, University of Southern California.
- Lin, J. (2006). The Role of Information Retrieval in Answering Complex Questions. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (COLING/ACL 2006)*, pp. 523-530. Sydney, Australia.
- Lin, J. et al. (2002). Extracting answers from the Web using knowledge annotation and knowledge mining techniques. In *Proceedings of the Eleventh Text Retrieval Conference (TREC-11)*.
- Linden, K. V. (1994). Generating precondition expressions in instructional text. In *Proceedings of the 32nd conference on Association for Computational Linguistics*, pp. 42-49. Las Cruces, New Mexico.
- Linden, K. V. and J. H. Martin (1995). Expressing rhetorical relations in instructional text: a case study of the purpose relation. *Computational Linguistics*, v.21 n.1, pp. 29-57.

References

- Lykke, M and A. Eslau (2010). Using thesauri in enterprise settings: indexing or query expansion? In: *The Janus Faced Scholar: a Festschrift in honour of Peter Ingwersen*. Det Informationsvidenskabelige Akademi, 2010. P. 87-97.
- Mani, I. (2001). *Automatic Summarization*. Amsterdam/Philadelphia: John Benjamins Publishing Company.
- Mani, I. and M. T. Maybury (1999). *Advances in Automatic Text Summarization*. MIT Press, 1999.
- Mann, W. C. and S. A. Thompson (1987). *Rhetorical Structure Theory: A Theory of Text Organization*. ISI: Information Sciences Institute, Los Angeles, CA, ISI/RS-87-190.
- Manning, C. and H. Schütze (1999). *Foundations of Statistical Natural Language Processing Cambridge*. Massachusetts: MIT Press.
- Manning, C., R. Prabhakar and H. Schütze (2008). *Introduction to Information Retrieval*. Cambridge University Press, Cambridge, 2008.
- Marcu, D. (1997). *The Rhetorical Parsing, Summarization, and Generation of Natural Language Texts*. Ph.D. dissertation. University of Toronto.
- Marcu, D. (2000). *The Theory and Practice of Discourse Parsing and Summarization*. Cambridge, MA: The MIT Press.
- Mathesius, V. (1915). O passivu v moderní angličtině. *Sborník filologický* 5, pp. 198-220.
- McCandless M., E. Hatcher and O. Gospodnetic (2010). *Lucene in Action* (2nd ed.). Manning Publications.
- McKeown, K. R. (1985). Discourse strategies for generating natural-language text. In B. L. Webber, B. Grosz, K. S. Jones, editor, *Readings in Natural Language Processing*. Morgan Kaufmann Publishers, 1986.
- McKeown, K.R. and D.R. Radev (1995). Generating Summaries of Multiple News Articles. In *Proceedings of the Eighteenth Annual International ACM Conference on Research and Development in Information Retrieval (SIGIR)*, pp. 74–82. Seattle, WA, USA.
- Mishne, G. and M. de Rijke (2005). Boosting Web Retrieval through Query Operations. In *27th European Conference on Information Retrieval (ECIR'05)*.

References

- Mitra, M., C. Buckley, A. Singhal and C. Cardie (1997). An analysis of statistical and syntactic phrases. In *Proceedings of RIAO-97, 5th International Conference on Recherche d'Information Assistee par Ordinateur*, pp. 200-214. Montreal, CA.
- Monz, C. (2003). *From Document Retrieval to Question Answering*. Ph.D. Thesis, Institute for Logic, Language and Computation, Universiteit van Amsterdam.
- Mooney, R. J. (1990). Learning plan schemata from observation: Explanation-based learning for plan recognition. *Cognitive Science*, 14(4), 483-509.
- Morris, J. and G. Hirst (1991). Lexical Cohesion Computed by Thesaural Relations as an Indicator of the Structure of Text. *Computational Linguistics* 17(1), pp. 21-48.
- Murdok, V. and B. Croft (2002). Task orientation in question answering. In *Proceedings of SIGIR '02*, pp. 355-356. Tampere, Finland.
- Neumann, G. and B. Sacaleanu (2003). A cross-language question/answering-system for German and English. In *Proceedings of The Working Notes for the CLEF 2003 Workshop*. Trondheim, Norway.
- Neumann, G., F. Xu and B. Sacaleanu (2003). Strategies for Web-based Cross-Language Question Answering. In *Proceedings of 2nd CoLogNET-ElsNET Symposium on Questions and Answers: Theoretical and Applied Perspectives*, pp. 84-95. Amsterdam.
- Norris, J. (1998). *The Generation of Compound Nominals to Represent the Essence of Text: The COMMIX System*. Ph.D. Thesis, University of Brighton.
- Noy, N. F. (1997). *Knowledge Representation for Intelligent Information Retrieval in Experimental Sciences*. Ph.D. Thesis, Northeastern University.
- Oates, S. (2001). *Generating Multiple Discourse Markers in Text*. M.Phil. Thesis, ITRI, University of Brighton.
- Ounis, I., G. Amati, V. Plachouras, B. He, C. Macdonald and C. Lioma (2006). Terrier: A high performance and scalable information retrieval platform. In: *Proc. of the SIGIR Workshop on Open Source Information Retrieval*.
- Paice, C.D. and P. A. Jones (1993). The Identification of important concepts in highly structured technical papers. In *Proceedings of the 16th ACM-SIGIR Conference*. Pittsburgh, PA.
- Paraboni, I. (2003). *Generating References in Hierarchical Domains: the Case of Document Deixis*. Ph.D. Thesis, University of Brighton.

References

- Parouty, M. (1993). *Mozart, from Child Prodigy to Tragic Hero*. New York: Harry N. Abrams.
- Partee, B. H. (1996). Allegation and local accommodation. In B. H. Partee and P. Sgall (Eds.), *Discourse and meaning: papers in honor of Eva Hajičová*, pp. 65–86. John Benjamins.
- Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible inference*. Morgan Kaufmann Publishers, Inc., 1988.
- Perfetti, C. A. and S. R. Goldman (1974). Thematization and sentence retrieval. *Journal of Verbal Learning and Verbal Behavior* 13, pp. 97-116.
- Picard, J. (1999). Finding content-bearing terms using term similarities. In *Proceedings of the Ninth Conference of the European Chapter of the Association for Computational Linguistics (EACL 1999)*, pp. 241-244. Bergen, Norway.
- Pickens, J. and W. Croft (2000). An exploratory analysis of phrases in text retrieval. In *Proceedings of RIAO2000*.
- Plachouras, V., B. He and I. Ounis (2004). University of Glasgow at TREC2004: Experiments in web, robust and terabyte tracks with Terrier. In *Proceedings of the Thirteenth Text REtrieval Conference (TREC 2004)*.
- Plante, P. (1980). *Une grammaire Déredéc des structures de surface du français appliquée à l'analyse de contenu de texts*. Montréal, Service de l'informatique, Université du Québec à Montréal.
- Ponte, J. and W. Croft (1998). A Language Model Approach to Information Retrieval. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 275–281.
- Power, R. (2004). *Extending the Notion of Topic*. Unpublished Report. ITRI, University of Brighton
- Prager, J. and J. Chu-Carroll (2001). Use of WordNet Hypernyms for Answering What-Is Questions. In *Proceedings of the TREC-10 Conference*, pp. 309-316. NIST, Gaithersburg, MD.
- Prince, E. F. (1992). The ZPG letter: Subjects, definiteness, and information status. In W. C. Mann & S. A. Thompson (eds.), *Discourse description: Diverse linguistic analyses of a fund-raising text*, pp. 295-326. Philadelphia: John Benjamins.
- Ravichandran, D. and E. H. Hovy (2002). Learning Surface Text Patterns for a Question Answering System. In *Proceedings of the ACL Conference*.

References

- Reinhart, T. (1982). Pragmatics and linguistics: An analysis of sentence topics. *Philosophica* 27, pp. 53-94.
- Reiter, E. and R. Dale (1997). Building Applied Natural-Language Generation Systems. *Journal of Natural-Language Engineering*, 3, pp. 57-87.
- Rino, L. H. M. and D. Scott (1996). A discourse model for gist preservation. In: D. L. Borges and C. A. A. Kaestner (eds.), *Advances in Artificial Intelligence. Lecture Notes in Artificial Intelligence*, Vol, 1159, pp. 131-140.
- Robertson, S.E. and S. Walker (1997). Some simple effective approximations to the 2-Poisson model for probabilistic weighted retrieval. In *Readings in information retrieval*. Morgan Kaufmann Publishers Inc. San Francisco, CA, USA.
- Robertson, S. E. and K. Sparck Jones (1976). Relevance weighting of search terms. *Journal of the American Society of Information Science*, 27, pp. 129-146.
- Robertson, S., S. Walker, S. Jones, M. Hancock-Beaulieu and M. Gatford (1993). Okapi at TREC-2. In *Proceedings of the Second Text Retrieval Conference (TREC-2)*, pp. 21-24.
- Robertson, S. E., S. Walker, S. Jones, M. Hancock-Beaulieu and M. Gatford (1996). Okapi at trec-3. In *Proceedings of the Third Text Retrieval Conference*, pp. 109-126.
- Rocchio, J. (1971). Relevance feedback in information retrieval. In G. Salton (Eds.) *The Smart Retrieval system -Experiments in Automatic Document Processing*, pp. 313-323. Prentice-Hall, Englewood Cliffs, NJ.
- Salkie, R. (2004). *Descriptive and Informative Abstracts and Summaries*. Handout of Module RMM06, CRM/DRM, University of Brighton.
- Salton, G., C. Buckley and M. Smith (1990). On the application of syntactic methodologies in automatic text analysis. *Information Processing and Management*, 26, 1990, 73-92.
- Salton, G., E. A. Fox and H. Wu (1983). Extended Boolean information retrieval. *Communications of the ACM*, 26(11), pp. 1022-1036.
- Salton, G. and M. J. McGill (Eds.) (1983). *Introduction to Modern Information Retrieval*. McGraw-Hill.
- Salton, G. and M. Smith (1989). On the application of syntactic methodologies in automatic text analysis. In *Twelfth Annual International ACM SIGIR Conference on Research and Development in information Retrieval*, pp. 137-150.

References

- Sanders, T. J. M., W. P. M. Spooren and L. G. M. Noordman (1992). Towards a taxonomy of coherence relations. *Discourse Processes*, 15, pp. 1-35.
- Sanderson, M. (1994). Word Sense Disambiguation and Information Retrieval. In *Proceedings of the 17th International ACM SIGIR*, pp. 49 – 57. Dublin, Ireland.
- Sanderson, M. (1997). *Word Sense Disambiguation and Information Retrieval*. Ph.D. Thesis, Technical Report (TR-1997-7) of the Department of Computing Science at the University of Glasgow, Glasgow, UK.
- Sanderson, M. (2000). Retrieving with good sense. *Information Retrieval*, 2 (1), pp. 49-69. ISSN 1573-7659.
- Sanderson, M. (2008). Ambiguous queries: test collections need more sense. *SIGIR 2008*, pp. 499-506.
- Santini, M. (2004). A Shallow Approach to Syntactic Feature Extraction for Genre Classification. In *Proceedings of the 7th Annual Colloquium for the UK Special Interest Group for Computational Linguistics (CLUK-04)*.
- Saxena, A. et al. (2007). IITD-IBMIRL system for question answering using pattern matching, semantic type and semantic category recognition. In *Proceedings of the Sixteenth Text Retrieval Conference (TREC 2007)*.
- Schank, R.C. and R.P. Abelson (1977). *Scripts, Plans, Goals, and Understanding*. Lawrence Erlbaum Associates, Hillsdale, NJ, USA.
- Schütze, H. and J. O. Pedersen (1995). Information retrieval based on word senses. In *Proceedings of the Symposium on Document Analysis and Information Retrieval*, 4, pp. 161-175.
- Schwitler, R., F. Rinaldi and S. Clematide (2004). The Importance of How-Questions in Technical Domains. In *Proceedings of the Question-Answering workshop of TALN 04*. Fez, Morocco.
- Scott, D (1996). Computer support for authoring multilingual software documentation. In *Proceedings of Translation and the Computer*. London.
- Sebastiani, F. (1999). Machine Learning in Automated Text Categorisation. *ACM Computing Surveys*, 34(1), pp. 1-47.

References

- Sebastiani, F. (2005). *Machine Learning*. Invited Lecture in the Fifth European Summer School on Information Retrieval (ESSIR'05). Dublin, Ireland.
- Sgall P., E. Hajičová and J. Panevová (1986). *The meaning of the sentence in its semantic and pragmatic aspects*. Dordrecht: Reidel - Prague: Academia.
- Shah, U., T. Finin, A. Joshi, J. Mayfield and R. Scott (2002). Information retrieval on the semantic web. In *Proceedings of the ACM Conference on Information and Knowledge Management*.
- Simmons, R. F. (1965). Answering English Questions by Computer: A Survey. *Communications Association for Computing Machinery (ACM)*, 8(1), pp. 53-70.
- Song, Y.-I., J.-T. Lee and H.-C. Rim (2009). Word or phrase?: learning which unit to stress for information retrieval. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pp. 1048 -1056.
- Soricut, R. and E. Brill (2004). Automatic Question Answering: Beyond the Factoid. In *Proceedings of the Human Language Technology and North American Association for Computational Linguistics Conference (HLT/NAACL-2004)*. Boston, MA.
- Spangler, S, J. T. Kreulen and J. Lessler (2002). Mindmap: Utilizing Multiple Taxonomies and Visualization to Understand a Document Collection. *HICSS 2002*, pp. 102.
- Sparck Jones, K. (2003). Document Retrieval: Shallow Data, Deep Theories; Historical Reflections, Potential Directions. *ECIR 2003*, pp. 1-11.
- Sparck Jones, K. and S. Robertson (2001). LM vs PM: where's the relevance? *Presentations at Workshop on Language Modelling and Information Retrieval*.
- Sparck Jones, K., S. E. Robertson and M. Sanderson (2007). Ambiguous requests: implications for retrieval tests, systems and theories. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in IR*, pp. 23-27.
- Spitters, M. and W. Kraaij (2001). TNO at TDT2001: Language model-based topic detection. In *Topic Detection and Tracking (TDT) Workshop 2001*. Gaithersburg, USA.
- Srihari, R., C. Niu and W. Li (2000). A hybrid approach for named entity and sub-type tagging. In *Proceedings of the sixth conference on Applied natural language processing*, pp. 247-254. Seattle, Washington.

References

- Stamatatos, E., N. Fakotakis and G. Kokkinakis (2000). Automatic text categorization in terms of genre and author. *Computational Linguistics*, 26(4), pp. 471-495.
- Stokoe, C., M. P. Oakes and J. Tait (2003). Word sense disambiguation in information retrieval revisited. In: *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval 2003*, pp. 159-166.
- Stricker, M., F. Vichot, G. Dreyfus and F. Wolinski (1999). Two-Step Feature Selection and Neural Network Classification for the TREC-8 Routing. In *the 8th Text Retrieval Conference (TREC-8)*. Gaithersburg, USA.
- Svore K., P. Kanami and N. Khan (2010). How good is a span of terms? Exploiting proximity to improve web retrieval. In *the Proceedings of SIGIR 2010*.
- Takechi, M., T. Tokunaga, Y. Matsumoto and H. Tanaka (2003). Feature selection in categorizing procedural expressions. In *Proceedings of the Sixth International Workshop on Information Retrieval with Asian Languages (IRAL2003)*, pp. 49-56. Sapporo, Japan.
- Todisrascu, A. (2001). *Semantic Indexing for Information Retrieval Systems*. Ph.D. thesis, ENSAIS, Strasbourg, France.
- Tsur, O., M. de Rijke and K. Sima'an (2004). BioGrapher: biography questions as a restricted domain question answering task. In *Proceedings ACL 2004 Workshop on Question Answering in Restricted Domains*. Barcelona.
- Turtle, H. and W. B. Croft (1991). Evaluation of an inference network based retrieval model. *Trans. Inf. Syst.*, 9(3), pp. 187-222.
- Turtle, H. R. and W. B. Croft (1992). A comparison of text retrieval models. *The Computer Journal* 35 (3), pp. 279–290.
- Vallduví, E. (1990). *The Informational Component*. Ph.D. dissertation. University of Pennsylvania.
- van Dijk, T. A. (1977). *Text and Context*. London: Longman.
- van Dijk, T. A. (1985). *Handbook of Discourse Analysis*. Vol. 2. Academic Press.
- van Rijsbergen, C. J. (1979). *Information Retrieval*, second edition. Butterworths, London, 1979. Accessed at <<http://www.dcs.gla.ac.uk/Keith/Preface.html>>

References

- Volk, M., S. Vintar and P. Buitelaar (2003). Ontologies in Cross-Language Information Retrieval. In: *Proceedings of WOW2003 (Workshop Ontologie-basieres Wissensmanagement)*. Luzern, Switzerland.
- Voorhees, E. M. (1993). Using WordNet to disambiguate word senses for text retrieval. In *Proceedings of ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 171-180.
- Voorhees, E. M. (2003). Overview of the TREC 2003 Question Answering Track. In *Proceeding of the Twelfth Text REtrieval Conference*.
- Voorhees, E. M. and D. Harman (2000). Overview of the eighth Text REtrieval Conference (TREC-8). In E. M. Voorhees and D. K. Harman (Eds.), *Proceedings of the Eighth Text Retrieval Conference (TREC-8)*, pp. 1-24. NIST Special Publication 500-246. Accessed at <<http://trec.nist.gov/pubs.html>>
- Weisstein, E. (1999). *Correlation Coefficient*. From MathWorld--A Wolfram Web Resource. <<http://mathworld.wolfram.com/CorrelationCoefficient.html>> [Accessed 21 Oct 2005]
- Wilbur, W.J. (2002). A Thematic Analysis of the AIDS Literature. *Pacific Symposium on Biocomputing 7*, pp. 386-397.
- Winograd, T. (1972). *Understanding Natural Language*. Academic Press, New York.
- Witten, I. H. and E. Frank (2000). *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann, San Mateo, CA.
- Xu, J., A. Licuanan and R. Weischedel (2003). TREC2003 QA at BBN: answering definitional questions. In *Proceedings of the Twelfth Text Retrieval Conference (TREC 2003)*, pp. 98-106.
- Yang, Y. (1999). An Evaluation of Statistical Approaches to Text Categorization. *Journal of Information Retrieval 1(1/2)*, pp. 67-88.
- Yang, Y. and X. Liu (1999). A Re-Examination of Text Categorization Methods. In *Proceedings of ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'99)*, pp. 42-49.
- Yin, L. (2004). *Topic analysis and answering procedural questions*. ITRI Technical Report 04-14, ITRI, University of Brighton. <<http://www.itri.brighton.ac.uk/techindex.html>> [accessed June 1st, 2005]

References

- Yin, L. (2006). A Two-Stage Approach to Retrieve Answers for How-To Questions. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*. Student session, Trento, Italy.
- Yin, L. and R. Power (2005). Investigating the structure of topic expressions: a corpus-based approach. In *Proceedings from the Corpus Linguistics Conference Series*, Vol.1, No.1, University of Birmingham, Birmingham.
- Yin, L. and R. Power (2006). Adapting the Naive Bayes classifier to rank procedural texts. In *Proceedings of the 28th European Conference on IR Research (ECIR 2006)*.
- Young, R. M. and J. D. Moore (1994). DPOCL: A Principled Approach to Discourse Planning. In *7th International Workshop on Natural Language Generation*, Kennebunkport, Maine.
- Zhang, D. (2004). Web based question answering with aggregation strategy. In *Proceedings of The 6th Asia Pacific Web Conference (APWEB2004)*.
- Zheng, Z. (2002). AnswerBus question answering system. In *Proceeding of HLT Human Language Technology Conference (HLT 2002)*, pp. 24 – 27.

APPENDIX A

WH-QUESTION ANALYSIS

For each WH-question, we recast it into the form of ‘What is/are the *G* of/for/that *S*’ to extract the generic part of the question. We include the results from two annotators and compare their results to see whether they derive the same generic part. More details are introduced in section 4.1.

A.1 QUESTIONS FROM MEDICAL DOMAIN

This question set contains 77 frequent asked questions in the medical domain.

<i>Question:</i> How much water do we need on a daily basis?
<i>Annotator1:</i> AMOUNT OF WATER
<i>Annotator2:</i> What is the amount of water that we need on a daily basis.
<i>Result analysis:</i> match

<i>Question:</i> Why should we drink so much water?
<i>Annotator1:</i> REASON FOR DRINKING
<i>Annotator2:</i> What is the reason that we need so much water?
<i>Result analysis:</i> match

<i>Question:</i> What causes humans to become thirsty?
<i>Annotator1:</i> CAUSE OF THIRST
<i>Annotator2:</i> What is the reason that humans become thirsty?
<i>Result analysis:</i> not match

<i>Question:</i> Why should we not drink sea-water?
<i>Annotator1:</i> REASON FOR NOT DRINKING SEA-WATER
<i>Annotator2:</i> What is the reason we should not drink sea-water?
<i>Result analysis:</i> match

<i>Question:</i> What health effects do microorganisms in water cause?
<i>Annotator1:</i> EFFECTS OF MICRO-ORGANISMS

Appendix A – W-H Question Analysis

<i>Annotator2</i> : What are the health effects that microorganisms cause?
<i>Result analysis</i> : match

<i>Question</i> : What is Cryptosporidium?
<i>Annotator1</i> : DEFINITION OF CRYPTOSPORIDIUM
<i>Annotator2</i> : What is the definition of Cryptosporidium?
<i>Result analysis</i> : match

<i>Question</i> : What are the risks of hard water?
<i>Annotator1</i> : RISKS OF HARD WATER
<i>Annotator2</i> : What are the risks of hard water?
<i>Result analysis</i> : match

<i>Question</i> : How soon will I see a change of my breathing when on Tracleer?
<i>Annotator1</i> : TIME UNTIL CHANGE OF BREATHING
<i>Annotator2</i> : What is the amount of time before I see a change when on Tracleer?
<i>Result analysis</i> : not match

<i>Question</i> : What are the side-effects of Tracleer?
<i>Annotator1</i> : EFFECTS OF TRACLEER
<i>Annotator2</i> : What are the side-effects of Tracleer?
<i>Result analysis</i> : match

<i>Question</i> : How long does it take Tracleer to work?
<i>Annotator1</i> : TIME UNTIL TRACLEER WORKS
<i>Annotator2</i> : What is the amount of time for Tracleer to take effect?
<i>Result analysis</i> : not match

<i>Question</i> : What have been the results of clinical trials?
<i>Annotator1</i> : RESULTS OF TRIALS
<i>Annotator2</i> : What are the results of clinical trials?
<i>Result analysis</i> : match

Appendix A – W-H Question Analysis

<u>Question:</u> How long can Tracleer treat a PH patient before benefits emerge?
<u>Annotator1:</u> TIME UNTIL BENEFITS
<u>Annotator2:</u> What is the length of time that Tracleer can treat a PH patient before benefits emerge?
<u>Result analysis:</u> not match

<u>Question:</u> What are the results of Tracleer being tested on children?
<u>Annotator1:</u> RESULTS OF TESTS
<u>Annotator2:</u> What are the results of Tracleer that are being tested on children?
<u>Result analysis:</u> match

<u>Question:</u> Why would you get tiny blisters on your tongue?
<u>Annotator1:</u> REASON FOR BLISTERS
<u>Annotator2:</u> What is the reason that you get tiny blisters on your tongue?
<u>Result analysis:</u> match

<u>Question:</u> When do you use rubbing alcohol?
<u>Annotator1:</u> CONDITIONS FOR RUBBING ALCOHOL
<u>Annotator2:</u> What is a symptom that would warrant the use of rubbing alcohol?
<u>Result analysis:</u> not match

<u>Question:</u> How much does a DNA test cost?
<u>Annotator1:</u> COST OF DNA TEST
<u>Annotator2:</u> What is the cost of a DNA test?
<u>Result analysis:</u> match

<u>Question:</u> How long after being exposed to a cold virus before symptoms appear?
<u>Annotator1:</u> TIME UNTIL SYMPTOMS
<u>Annotator2:</u> What is the time period that a symptoms appear after being exposed to a cold virus?
<u>Result analysis:</u> not match

<u>Question:</u> What are the side-effects of protein supplements?
<u>Annotator1:</u> EFFECTS OF SUPPLEMENTS
<u>Annotator2:</u> What are the side-effects of protein supplements?
<u>Result analysis:</u> match

Appendix A – W-H Question Analysis

<i>Question:</i> Why would it be blue under your eyes?
<i>Annotator1:</i> CAUSE OF BLUE UNDER EYES
<i>Annotator2:</i> What would be the cause of the blue color under your eyes?
<i>Result analysis:</i> match

<i>Question:</i> What causes Chlamydia?
<i>Annotator1:</i> CAUSE OF CHLAMYDIA
<i>Annotator2:</i> What is the REASON that Chlamydia is contracted?
<i>Result analysis:</i> not match

<i>Question:</i> How do you cure a child's bad nose bleed?
<i>Annotator1:</i> CURE FOR NOSE BLEED
<i>Annotator2:</i> What is the cure for a child's bad nose bleed?
<i>Result analysis:</i> match

<i>Question:</i> What causes a hump in the back ...
<i>Annotator1:</i> CAUSE OF HUMP
<i>Annotator2:</i> What is the reason that a hump develops in the back?
<i>Result analysis:</i> not match

<i>Question:</i> What can be done about it?
<i>Annotator1:</i> CURE/TREATMENT FOR HUMP
<i>Annotator2:</i> What are the possibilities of how to handle it?
<i>Result analysis:</i> not match

<i>Question:</i> What is life like for an autistic adult?
<i>Annotator1:</i> EXPERIENCE OF AUTISTIC ADULT
<i>Annotator2:</i> What is life like for an autistic adult?
<i>Result analysis:</i> Annotator2's result does not match the requirement

<i>Question:</i> Why would you get headaches after a nose bleed?
<i>Annotator1:</i> CAUSE OF HEADACHES
<i>Annotator2:</i> What is the cause of headaches after a nose bleed?
<i>Result analysis:</i> match

<i>Question:</i> What type of supplement can a man take to make his hair grow?
<i>Annotator1:</i> TYPE OF SUPPLEMENT
<i>Annotator2:</i> What is the type of supplement that a man can take to make his hair grow?
<i>Result analysis:</i> match

Appendix A – W-H Question Analysis

<i>Question:</i> How can you cope with fibromyalgia?
<i>Annotator1:</i> METHOD OF COPING / TREATMENT FOR FIBROMYALGIA
<i>Annotator2:</i> What is the best way to cope with fibromyalgia?
<i>Result analysis:</i> not match

<i>Question:</i> What are the effects of eating brown lettuce?
<i>Annotator1:</i> EFFECTS OF EATING
<i>Annotator2:</i> What are the effects of eating brown lettuce?
<i>Result analysis:</i> match

<i>Question:</i> What colour should semen be?
<i>Annotator1:</i> COLOUR OF SEMEN
<i>Annotator2:</i> What is the actual colour of semen?
<i>Result analysis:</i> match

<i>Question:</i> What parts of the body are most vulnerable to lightning damage?
<i>Annotator1:</i> LOCUS OF DAMAGE
<i>Annotator2:</i> What are the parts of the body that are most vulnerable to lightning damage?
<i>Result analysis:</i> not match

<i>Question:</i> Why does the gland below the jaw swell when eating solid food?
<i>Annotator1:</i> CAUSE OF SWELLING
<i>Annotator2:</i> What is the cause of swelling of the gland below the jaw when eating solid food?
<i>Result analysis:</i> match

<i>Question:</i> What could cause pain throughout the body?
<i>Annotator1:</i> CAUSE OF PAIN
<i>Annotator2:</i> What is the reason that a pain would radiate throughout the body?
<i>Result analysis:</i> not match

<i>Question:</i> What is the diagnosis/remedy for someone who falls asleep whenever they sit?
<i>Annotator1:</i> DIAGNOSIS OF SOMEONE/ REMEDY FOR FALLING ASLEEP
<i>Annotator2:</i> What is the diagnosis/remedy for someone who falls asleep whenever they sit?
<i>Result analysis:</i> match

Appendix A – W-H Question Analysis

<u>Question</u> : How big is the average penis?
<u>Annotator1</u> : SIZE OF PENIS
<u>Annotator2</u> : What is the average size of a penis?
<u>Result analysis</u> : match

<u>Question</u> : What would cause your eyes to be slightly puffy, red, and seeping fluid?
<u>Annotator1</u> : CAUSE OF PUFFY EYES
<u>Annotator2</u> : What are the causes for your eyes to be slightly puffy, red, and seeping fluid?
<u>Result analysis</u> : match

<u>Question</u> : If you experience twitching in your fingers ..., what could this be?
<u>Annotator1</u> : EXPLANATION OF TWITCHING
<u>Annotator2</u> : What is the reason that you may experience twitching in your fingers?
<u>Result analysis</u> : not match

<u>Question</u> : What kind of liquid is in instant cold packs?
<u>Annotator1</u> : TYPE OF LIQUID
<u>Annotator2</u> : What is the type of liquid in instant cold packs?
<u>Result analysis</u> : match

<u>Question</u> : What is the effect of Ige?
<u>Annotator1</u> : EFFECT OF IGE
<u>Annotator2</u> : What is the effect of Ige?
<u>Result analysis</u> : match

<u>Question</u> : What is a rash in the shape of a circle on your skin?
<u>Annotator1</u> : TYPE OF RASH
<u>Annotator2</u> : What is a rash in the shape of a circle on your skin?
<u>Result analysis</u> : Annotator2's result does not match the requirement

<u>Question</u> : Where can you use rubbing alcohol?
<u>Annotator1</u> : LOCUS FOR RUBBING ALCOHOL
<u>Annotator2</u> : What is the area where you can use rubbing alcohol?
<u>Result analysis</u> : not match

<u>Question</u> : What does "beauty is only skin deep" mean?
<u>Annotator1</u> : MEANING OF SENTENCE
<u>Annotator2</u> : What is the meaning of "beauty is only skin deep?"

Appendix A – W-H Question Analysis

Result analysis: match

Question: If you have stopped having bowel movements ..., what could be wrong with you?

Annotator1: EXPLANATION FOR NO BOWEL MOVEMENTS

Annotator2: What is potentially wrong with you if you have stopped having bowel movements?

Result analysis: Annotator2's result does not match the requirement

Question: What can one expect if a biopsy comes back showing a basal cell skin cancer?

Annotator1: INTERPRETATION OF BIOPSY

Annotator2: What is the possible outcome if the diagnosis is a basal cell skin cancer?

Result analysis: not match

Question: What does rubbing alcohol have in it?

Annotator1: COMPOSITION OF RUBBING ALCOHOL

Annotator2: What are the ingredients in rubbing alcohol?

Result analysis: not match

Question: What is the definition of mental illness?

Annotator1: DEFINITION OF MENTAL ILLNESS

Annotator2: What is the definition of mental illness?

Result analysis: match

Question: What is the treatment of mental illness?

Annotator1: TREATMENT OF MENTAL ILLNESS

Annotator2: What is the treatment of mental illness?

Result analysis: match

Question: What causes mental illness?

Annotator1: CAUSE OF MENTAL ILLNESS

Annotator2: What are the causes of mental illness?

Result analysis: match

Question: Who gets mental illness?

Annotator1: TYPE OF SUFFERER

Annotator2: What are the types of people that acquire mental illness?

Result analysis: match

Question: What is a psychiatrist/psychologist?

Appendix A – W-H Question Analysis

<u>Annotator1</u> : DEFINITION OF PSYCHIATRIST
<u>Annotator2</u> : What is a psychiatrist/psychologist?
<u>Result analysis</u> : Annotator2's result does not match the requirement

<u>Question</u> : What happens in a psychiatric hospital?
<u>Annotator1</u> : DISTINCTIVE OCCURRENCES IN HOSPITAL
<u>Annotator2</u> : What are the kinds of things that occur in a psychiatric hospital?
<u>Result analysis</u> : not match

<u>Question</u> : Who takes medication?
<u>Annotator1</u> : RECIPIENT OF MEDICATION
<u>Annotator2</u> : What are the types of people that take medication?
<u>Result analysis</u> : not match

<u>Question</u> : What are the side-effects of medication?
<u>Annotator1</u> : EFFECTS OF MEDICATION
<u>Annotator2</u> : What are the side-effects of medication?
<u>Result analysis</u> : match

<u>Question</u> : Why do people commit suicide?
<u>Annotator1</u> : CAUSE/MOTIVE FOR SUICIDE
<u>Annotator2</u> : What is the reason that people commit suicide?
<u>Result analysis</u> : not match

<u>Question</u> : How do you know if someone is suicidal?
<u>Annotator1</u> : SYMPTOMS (WARNING SIGNS) OF SUICIDE
<u>Annotator2</u> : What is a way to tell if someone is suicidal?
<u>Result analysis</u> : not match

<u>Question</u> : What should you do if a friend is thinking of suicide?/What do you do in an emergency or suicide attempt?
<u>Annotator1</u> : APPROPRIATE RESPONSE TO SUICIDE DANGER/ATTEMPT
<u>Annotator2</u> : What is the right thing to do if a friend is thinking of suicide?/What are methods of handling an emergency or a suicide attempt?
<u>Result analysis</u> : not match

<u>Question</u> : What is a community mental health centre?
<u>Annotator1</u> : DESCRIPTION OF MENTAL HEALTH CENTRE
<u>Annotator2</u> : What is a community mental health center?
<u>Result analysis</u> : Annotator2's result does not match the requirement

Appendix A – W-H Question Analysis

<i>Question:</i> Why would someone go to a CMHC?
<i>Annotator1:</i> REASON FOR ATTENDING CMHC
<i>Annotator2:</i> What is the reason someone goes to a CMHC?
<i>Result analysis:</i> match
<i>Question:</i> Why is antibiotic resistance important?
<i>Annotator1:</i> IMPORTANCE OF RESISTANCE
<i>Annotator2:</i> What is the importance to antibiotic resistance?
<i>Result analysis:</i> match
<i>Question:</i> What does antibiotic resistance mean?
<i>Annotator1:</i> MEANING OF ANTIBIOTIC RESISTANCE
<i>Annotator2:</i> What is the definition of antibiotic resistance?
<i>Result analysis:</i> not match
<i>Question:</i> What is the reason for antibiotic resistance to be a problem now?
<i>Annotator1:</i> EXPLANATION FOR PROBLEM
<i>Annotator2:</i> What is the reason for antibiotic resistance to be a problem now?
<i>Result analysis:</i> not match
<i>Question:</i> So what can we do (about antibiotic resistance)?
<i>Annotator1:</i> APPROPRIATE RESPONSE TO ANTIBIOTIC RESISTANCE
<i>Annotator2:</i> What is the best way to deal with antibiotic resistance?
<i>Result analysis:</i> not match
<i>Question:</i> How can we do it?
<i>Annotator1:</i> METHOD FOR DOING IT
<i>Annotator2:</i> What is a strategy?
<i>Result analysis:</i> Annotator2's result does not match the requirement
<i>Question:</i> How do I know whether it is a viral infection?
<i>Annotator1:</i> SYMPTOMS/EVIDENCE FOR VIRAL INFECTION
<i>Annotator2:</i> What is the best way to tell whether it is a viral infection?
<i>Result analysis:</i> not match
<i>Question:</i> How will I get better if antibiotics are not the answer?
<i>Annotator1:</i> TREATMENT/CURE FOR VIRAL INFECTION
<i>Annotator2:</i> What is another answer if I don't get better with antibiotics?
<i>Result analysis:</i> Annotator2's result does not match the requirement

Appendix A – W-H Question Analysis

<u>Question</u> : What should I do if my children keep getting infections?
<u>Annotator1</u> : (METHOD FOR) PREVENTION OF VIRAL INFECTIONS
<u>Annotator2</u> : What is a way to handle the situation if my children keep getting infections?
<u>Result analysis</u> : not match

<u>Question</u> : When are antibiotics the answer?
<u>Annotator1</u> : APPROPRIATE CONDITIONS FOR ANTIBIOTICS
<u>Annotator2</u> : What is the right sign that antibiotics are the answer?
<u>Result analysis</u> : not match

<u>Question</u> : When should I stop taking antibiotics?
<u>Annotator1</u> : APPROPRIATE DURATION OF ANTIBIOTICS TREATMENT
<u>Annotator2</u> : What is the time to know when I should stop taking antibiotics?
<u>Result analysis</u> : not match

<u>Question</u> : What is a health service provider required to do when asked to correct records?
<u>Annotator1</u> : APPROPRIATE RESPONSE TO REQUEST
<u>Annotator2</u> : What is a health service provider required to do when asked to correct records?
<u>Result analysis</u> : Annotator2's result does not match the requirement

<u>Question</u> : How should a request for access records be made?
<u>Annotator1</u> : PROCEDURE FOR ACCESS REQUEST
<u>Annotator2</u> : What is the best way that I should request access records?
<u>Result analysis</u> : not match

<u>Question</u> : How much time does an organisation have to meet an access request?
<u>Annotator1</u> : APPROPRIATE TIME UNTIL RESPONSE TO REQUEST
<u>Annotator2</u> : What is the amount of time it takes for an organization to have to meet an access request?
<u>Result analysis</u> : not match

<u>Question</u> : What is the relationship between professional confidentiality obligations ...
<u>Annotator1</u> : RELATIONSHIP BETWEEN PROFESSIONAL AND LEGAL OBLIGATIONS
<u>Annotator2</u> : What is the relationship between professional confidentiality obligations ...
<u>Result analysis</u> : match

<u>Question</u> : What privacy issues should be considered when the business circumstances ...
--

Appendix A – W-H Question Analysis

<i>Annotator1</i> : EFFECT OF BUSINESS CHANGE ON PRIVACY
<i>Annotator2</i> : What are the privacy issues that should be considered when the business circumstances ...
<i>Result analysis</i> : not match

<i>Question</i> : What are the restrictions on how the Medicare number can be handled ...
<i>Annotator1</i> : RESTRICTIONS ON USING MEDICARE NUMBER
<i>Annotator2</i> : What are the restrictions on how the Medicare number can be handled ...
<i>Result analysis</i> : match

<i>Question</i> : What should a health service provider tell an individual when ...
<i>Annotator1</i> : INSTRUCTIONS/EXPLANATIONS FOR INDIVIDUAL
<i>Annotator2</i> : What is the proper thing for a health service provider to tell an individual when ...
<i>Result analysis</i> : not match

<i>Question</i> : What should an organisation do with health information it no longer uses?
<i>Annotator1</i> : APPROPRIATE MANAGEMENT/HANDLING OF UNUSED INFORMATION
<i>Annotator2</i> : What is the right thing for an organization to do with health information that it no longer uses?
<i>Result analysis</i> : not match

<i>Question</i> : What privacy concerns should HSPs be aware of when ...
<i>Annotator1</i> : CONSTRAINTS ON PRIVACY
<i>Annotator2</i> : What are the privacy concerns that HSPs should be aware of when ...
<i>Result analysis</i> : not match

<i>Question</i> : Who are the guidelines on privacy ... for?
<i>Annotator1</i> : INTENDED USERS (APPLIERS) OF GUIDELINES
<i>Annotator2</i> : What are types of people who the guidelines on privacy are for?
<i>Result analysis</i> : not match

A.2 QUESTIONS FROM TREC 2004

This question set contains 60 questions used in the question answering track in TREC 2004.

Title: Crips

<i>Question</i> : When was the first Crip gang started?
<i>Annotator1</i> : What is the date of the start of the first Crip gang?
<i>Annotator2</i> : What date did the first Crip gang start?
<i>Result analysis</i> : match

Appendix A – W-H Question Analysis

<i>Question:</i> What does the name mean or come from?
<i>Annotator1:</i> What is the meaning or origin of the name?
<i>Annotator2:</i> What is the name's meaning or where does it come from?
<i>Result analysis:</i> match

<i>Question:</i> Which cities have Crip gangs?
<i>Annotator1:</i> What are the names of cities that have Crip gangs?
<i>Annotator2:</i> What particular cities have Crip gangs?
<i>Result analysis:</i> not match

<i>Question:</i> What ethnic group/race are Crip members?
<i>Annotator1:</i> What is the ethnic group/race of Crip members?
<i>Annotator2:</i> What is the ethnicity or race of Crip members?
<i>Result analysis:</i> match

<i>Question:</i> What is their gang color?
<i>Annotator1:</i> What is the color associated with their gang?
<i>Annotator2:</i> What is their gang color?
<i>Result analysis:</i> match

Title: Fred Durst

<i>Question:</i> What is the name of Durst's group?
<i>Annotator1:</i> What is the name of Durst's group?
<i>Annotator2:</i> What is Durst's group name?
<i>Result analysis:</i> match

<i>Question:</i> What record company is he with?
<i>Annotator1:</i> What is the name of the record company he is with?
<i>Annotator2:</i> What is the name of the record company he is with?
<i>Result analysis:</i> match

<i>Question:</i> What are titles of the group's releases?
<i>Annotator1:</i> What are titles of the group's releases?
<i>Annotator2:</i> What are the titles of the group's releases?
<i>Result analysis:</i> match

<i>Question:</i> Where was Durst born?
<i>Annotator1:</i> What is the place where Durst was born?
<i>Annotator2:</i> What is the place where Durst was born?

Appendix A – W-H Question Analysis

Result analysis: match

Title: Hale Bopp comet

Question: When was the comet discovered?

Annotator1: What is the date that the comet was discovered?

Annotator2: What is the date when the comet was discovered?

Result analysis: match

Question: How often does it approach the earth?

Annotator1: What is the frequency of its approach to the earth?

Annotator2: What is the frequency in which it approaches the earth?

Result analysis: match

Question: In what countries was the comet visible on its last return?

Annotator1: What are the names of the countries that the comet was visible in on its last return?

Annotator2: What are the countries where the comet was visible on its last return?

Result analysis: not match

Title: James Dean

Question: When was James Dean born?

Annotator1: What is the date that James Dean was born on?

Annotator2: What is the date when James Dean was born?

Result analysis: match

Question: When did James Dean die?

Annotator1: What is the date that James Dean died?

Annotator2: What is the date when James Dean died?

Result analysis: match

Question: How did he die?

Annotator1: What is the manner in which he died?

Annotator2: What is the way that he died?

Result analysis: not match

Question: What movies did he appear in?

Annotator1: What are the names of the movies he appeared in?

Annotator2: What are the movies that he appeared in?

Result analysis: not match

Appendix A – W-H Question Analysis

<i>Question:</i> Which was the first movie that he was in?
<i>Annotator1:</i> What is the name of the first movie he was in?
<i>Annotator2:</i> What is the first movie he was in?
<i>Result analysis:</i> not match

Title: AARP

<i>Question:</i> What does AARP stand for?
<i>Annotator1:</i> What is the expanded form of the abbreviation AARP?
<i>Annotator2:</i> What is the meaning of AARP?
<i>Result analysis:</i> not match

<i>Question:</i> When was the organization started?
<i>Annotator1:</i> What is the date on which the organization started?
<i>Annotator2:</i> What is the time when the organization started?
<i>Result analysis:</i> not match

<i>Question:</i> Where is its headquarters?
<i>Annotator1:</i> What is the location of its headquarters?
<i>Annotator2:</i> What is the location of its headquarters?
<i>Result analysis:</i> match

<i>Question:</i> Who is its top official or CEO?
<i>Annotator1:</i> What is the name of its top official or CEO?
<i>Annotator2:</i> What is the name of its top official or CEO?
<i>Result analysis:</i> match

<i>Question:</i> What companies has AARP endorsed?
<i>Annotator1:</i> What are the names of companies AARP has endorsed?
<i>Annotator2:</i> What are the names of the companies that AARP endorsed?
<i>Result analysis:</i> match

Title: Rhodes scholars

<i>Question:</i> How long does one study as a Rhodes scholar?
<i>Annotator1:</i> What is the duration of study as a Rhodes scholar?
<i>Annotator2:</i> What is the length of time which one studies as a Rhodes scholar?
<i>Result analysis:</i> not match

<i>Question:</i> Where do Rhodes scholars study?
<i>Annotator1:</i> What is the location of study of Rhodes Scholars?

Appendix A – W-H Question Analysis

<i>Annotator2</i> : What is the place where Rhodes scholars study?
<i>Result analysis</i> : not match

<i>Question</i> : What countries have Rhodes scholars come from?
<i>Annotator1</i> : What are the names of countries that Rhodes scholars come from?
<i>Annotator2</i> : What are the countries where Rhodes scholars come from?
<i>Result analysis</i> : not match

Title: agouti

<i>Question</i> : What kind of animal is an agouti?
<i>Annotator1</i> : What is the kind of the animal an agouti?
<i>Annotator2</i> : What animal is an agouti?
<i>Result analysis</i> : not match

<i>Question</i> : What is their average life span?
<i>Annotator1</i> : What is the length of their average life span?
<i>Annotator2</i> : What is their average life span?
<i>Result analysis</i> : not match

<i>Question</i> : In what countries are they found?
<i>Annotator1</i> : What are the names of the countries they are found in?
<i>Annotator2</i> : What are the countries where they are found?
<i>Result analysis</i> : not match

Title: Black Panthers

<i>Question</i> : Who founded the Black Panthers organization?
<i>Annotator1</i> : What is the name of the founder of the Black Panthers organization?
<i>Annotator2</i> : What is the name of the person who founded the Black Panthers organization?
<i>Result analysis</i> : match

<i>Question</i> : When was it founded?
<i>Annotator1</i> : What is the date of its founding?
<i>Annotator2</i> : What is the date when it was founded?
<i>Result analysis</i> : match

<i>Question</i> : Where was it founded?
<i>Annotator1</i> : What is the location of its founding?
<i>Annotator2</i> : What is the location where it was founded?
<i>Result analysis</i> : match

Appendix A – W-H Question Analysis

<i>Question:</i> Who have been members of the organization?
<i>Annotator1:</i> What are the names of people who have been members of the organization?
<i>Annotator2:</i> What are the names of those who have been members of the organization?
<i>Result analysis:</i> match

Title: Insane Clown Posse

<i>Question:</i> Who are the members of this group?
<i>Annotator1:</i> What are the names of the members of this group?
<i>Annotator2:</i> What are the names of the members of this group?
<i>Result analysis:</i> match

<i>Question:</i> What albums have they made?
<i>Annotator1:</i> What are the names of albums they have made?
<i>Annotator2:</i> What are the albums they made?
<i>Result analysis:</i> not match

<i>Question:</i> What is their style of music?
<i>Annotator1:</i> What is the style of their music?
<i>Annotator2:</i> What is their style of music?
<i>Result analysis:</i> match

<i>Question:</i> What is their biggest hit?
<i>Annotator1:</i> What is the title of their biggest hit?
<i>Annotator2:</i> What is their biggest hit?
<i>Result analysis:</i> not match

Title: Prions

<i>Question:</i> What are prions made of?
<i>Annotator1:</i> What is the composition of prions?
<i>Annotator2:</i> What are the elements contained within prions?
<i>Result analysis:</i> not match

<i>Question:</i> Who discovered prions?
<i>Annotator1:</i> What is the name of the discoverer of prions?
<i>Annotator2:</i> What are the names of those who discovered prions?
<i>Result analysis:</i> match

<i>Question:</i> What diseases are prions associated with?
--

Appendix A – W-H Question Analysis

<i>Annotator1</i> : What are the names of diseases that prions are associated with?
<i>Annotator2</i> : What are the diseases that are associated with prions?
<i>Result analysis</i> : not match

<i>Question</i> : What researchers have worked with prions?
<i>Annotator1</i> : What are the names of researchers who have worked on prions?
<i>Annotator2</i> : What are the names of the researchers who have worked with prions?
<i>Result analysis</i> : match

Title: the band Nirvana

<i>Question</i> : Who is the lead singer/musician in Nirvana?
<i>Annotator1</i> : What is the name of the lead singer/musician in Nirvana?
<i>Annotator2</i> : What is the name of the lead singer/musician in Nirvana?
<i>Result analysis</i> : match

<i>Question</i> : Who are the band members?
<i>Annotator1</i> : What are the names of the band members?
<i>Annotator2</i> : What are the names of the band members?
<i>Result analysis</i> : match

<i>Question</i> : When was the band formed?
<i>Annotator1</i> : What is the date of the band's formation?
<i>Annotator2</i> : What is the time when the band was formed?
<i>Result analysis</i> : not match

<i>Question</i> : What is their biggest hit?
<i>Annotator1</i> : What is the title of their biggest hit?
<i>Annotator2</i> : What is their biggest hit?
<i>Result analysis</i> : not match

<i>Question</i> : What are their albums?
<i>Annotator1</i> : What are the names of their albums?
<i>Annotator2</i> : What are the names of their albums?
<i>Result analysis</i> : match

<i>Question</i> : What style of music do they play?
<i>Annotator1</i> : What is the style of the music that they play?
<i>Annotator2</i> : What is the style of music they play?
<i>Result analysis</i> : match

Appendix A – W-H Question Analysis

Title: Rohm and Haas

<i>Question:</i> What industry is Rohm and Haas in?
<i>Annotator1:</i> What is the name of the industry that Rohm and Haas is in?
<i>Annotator2:</i> What is the industry that Rohm and Haas are in?
<i>Result analysis:</i> not match

<i>Question:</i> Where is the company located?
<i>Annotator1:</i> What is the location of the company?
<i>Annotator2:</i> What is the location of the company?
<i>Result analysis:</i> match

<i>Question:</i> What is their annual revenue?
<i>Annotator1:</i> What is the amount of their annual revenue?
<i>Annotator2:</i> What is their annual revenue?
<i>Result analysis:</i> not match

<i>Question:</i> How many employees does it have?
<i>Annotator1:</i> What is the number of employees that it has?
<i>Annotator2:</i> What is the number of employees it has?
<i>Result analysis:</i> match

Title: Jar Jar Binks

<i>Question:</i> What film introduced Jar Jar Binks?
<i>Annotator1:</i> What is the name of the film that introduced Jar Jar Binks?
<i>Annotator2:</i> What is the name of the film that introduced Jar Jar Binks?
<i>Result analysis:</i> match

<i>Question:</i> What actor is used as his voice?
<i>Annotator1:</i> What is the name of the actor who is used as his voice?
<i>Annotator2:</i> What is the name of the actor who used his voice?
<i>Result analysis:</i> match

<i>Question:</i> To what alien race does he belong?
<i>Annotator1:</i> What is the name of the alien race that he belongs to?
<i>Annotator2:</i> What is the alien race that he belongs to?
<i>Result analysis:</i> not match

Title: Horus

<i>Question:</i> Horus is the god of what?
<i>Annotator1:</i> What domain is Horus the god of?

Appendix A – W-H Question Analysis

<i>Annotator2</i> : What is Horus the god of?
<i>Result analysis</i> : not match

<i>Question</i> : What country is he associated with?
<i>Annotator1</i> : What is the name of the country he is associated with?
<i>Annotator2</i> : What is the country that he is associated with?
<i>Result analysis</i> : not match

<i>Question</i> : Who was his mother?
<i>Annotator1</i> : What is the name of his mother?
<i>Annotator2</i> : What is the name of his mother?
<i>Result analysis</i> : match

<i>Question</i> : Who was his father?
<i>Annotator1</i> : What is the name of his father?
<i>Annotator2</i> : What is the name of his father?
<i>Result analysis</i> : match

Title: Rat Pack

<i>Question</i> : Who are the members of the Rat Pack?
<i>Annotator1</i> : What are the names of the members of the Rat Pack?
<i>Annotator2</i> : What is the names of the Rat Pack members?
<i>Result analysis</i> : match

<i>Question</i> : Who coined the name?
<i>Annotator1</i> : What is the name of the person who coined the name?
<i>Annotator2</i> : What is the name of the person who coined the name?
<i>Result analysis</i> : match

<i>Question</i> : What Las Vegas hotel was made famous by the Rat Pack?
<i>Annotator1</i> : What is the name of the Las Vegas hotel which was made famous by the Rat Pack?
<i>Annotator2</i> : What is the name of the Las Vegas hotel that made the Rat Pack famous?
<i>Result analysis</i> : match

A.3 QUESTIONS FROM TREC 2007

This question set contains 5 questions from the ciQA task in TREC 2007.

<i>Question</i> : What is the position of Hank Aaron with respect to Barry Bonds' use of
--

Appendix A – W-H Question Analysis

steroids?
<u>Annotator1</u> : What is the position of Hank Aaron with respect to Barry Bonds' use of steroids?
<u>Annotator2</u> : With respect to Barry Bond's use of steroids, what is Hank Aaron's position?
<u>Result analysis</u> : match

<u>Question</u> : What evidence is there for transport of illegal immigrants from Croatia to the European Union?
<u>Annotator1</u> : What is the evidence that exists for transport of illegal immigrants from Croatia to the European Union?
<u>Annotator2</u> : What is the evidence for transporting illegal immigrants from Croatia to the European Union?
<u>Result analysis</u> : match

<u>Question</u> : What financial relationships exist between Google and its advertisers?
<u>Annotator1</u> : What is the financial relationship that exists between Google and its advertisers?
<u>Annotator2</u> : What is the financial relationship that exists between Google and its advertisers?
<u>Result analysis</u> : match

<u>Question</u> : What common interests exist between President Bush and Bono, the U2 Rock Star?
<u>Annotator1</u> : What are the interests that President Bush and Bono, the U2 Rock Star have in common?
<u>Annotator2</u> : What are the common interests that exist between President Bush and Bono, the U2 Rock Star?
<u>Result analysis</u> : match

<u>Question</u> : What effect does the Red Tide have on sea creatures?
<u>Annotator1</u> : What is the effect of the Red Tide on sea creatures?
<u>Annotator2</u> : What is the effect that the Red Tide has on sea creatures?
<u>Result analysis</u> : match

APPENDIX B

GENERIC CONCEPT AND PASSAGE SELECTION

In section 4.3, we introduce an experiment to verify that for a given generic concept, whether human subjects could reliably pick up passages that are relevant to it. This appendix contains the experiment materials used in that experiment. As mentioned in section 4.3, we prepare six passages for each generic concept, including three relevant ones and three irrelevant ones. The prediction of relevancy is done by the author based on the way she collected the passages and her own understanding of the content of the passages.

B.1 EXPERIMENT MATERIAL

GENERIC CONCEPT: CAUSE

Passage No. 1.a; Prediction: relevant
Title: What causes mental illness?
No-one knows all the reasons why people become mentally ill. Some people have a 'chemical imbalance' which affects how their brain works. This makes them have strange thoughts or feelings, or behave oddly. They may need to take medication to help their brain work better. For other people, something might happen in their life which is very stressful, such as a death of someone very close, and this may trigger a mental illness. Mental illness doesn't normally start out of the blue. It usually develops slowly. But some people do get a mental illness suddenly, such as when someone has a psychotic illness.

Passage No. 1.b; Prediction: relevant
Title: What causes depression?
Depression is triggered by a complex combination of genetic, psychological and environmental factors. Genetic means that in some families, depression is inherited, passed down through genes. Psychological makeup has to do with personality traits and environmental factors means life circumstances. The brain is an organ of the body just like the heart, liver and kidneys. If the chemicals in the brain (neurotransmitters) that regulate how a person thinks, feels and acts, get out of balance, the brain can get "sick" and the result can be clinical depression. A bad or stressful life event could trigger depression, however, a person can also be born with depression. It can also appear out of nowhere,

Appendix B – Generic Concept Analysis and Passage Selection

when everything is going fine, at a time when there is no reason to get depression. Depression is nothing to be ashamed of!

Passage No. 1.c; Prediction: relevant

Title: Why do people commit suicide?

Some suicides are the result of impulsive decisions based on a situation that seems hopeless - loss of a job, divorce, or a breakup with one's girlfriend or boyfriend. Suicide attempts triggered by major disappointments, such as romantic rejection, problems with peers, or failing a big exam, are common among depressed teenagers, who haven't had the life experience to realize that these "injuries" heal with time.

Ninety percent of the people who commit suicide have a mental or substance abuse disorder (or both). More than half of the people who kill themselves are seriously or clinically depressed.

Passage No. 1.d; Prediction: irrelevant

Title: What is clinical depression?

Depression is a medical illness, just like cancer or diabetes. It is not the "blues". The blues are normal feelings that eventually pass. The feelings associated with depression last longer than a couple of weeks. If your friend has depression, he can't talk himself out of it. Your friend isn't weak and doesn't have a character flaw. Having depression isn't his fault. Depression affects the whole body – thoughts, feelings, behavior, physical health, appearance, and all areas of a person's life – home, work, school and social life. Depression can be treated successfully just like other illnesses.

Passage No. 1.e; Prediction: irrelevant

Title: Does mental illness exist?

The English-speaking world has not always used medical language to describe the behavior we now label as symptomatic of mental illness or mental disorder. Descriptions were sometimes framed in quite different terms, such as possession. What we now call mental illness was not always treated as a medical problem. Non-English-speaking nations in the West have had changes in their linguistic usage and their treatment of the mentally ill roughly parallel to Anglophone countries. Anthropological work in non-Western cultures

Appendix B – Generic Concept Analysis and Passage Selection

suggests that there are many cases of behavior that psychiatry would classify as symptomatic of mental disorder, which are not seen within their own cultures as signs of mental illness. Indeed, other cultures may not even have a concept of mental illness that corresponds even approximately to the Western concept.

Passage No. 1.f; Prediction: irrelevant

Title: Can caviar cure depression?

It might sound a little fishy, but there is growing evidence that caviar can help chase away the blues. Early research suggests that people suffering from depression, bipolar disorder, and other mental health problems can benefit from diets rich in omega-3 fatty acids -- found in abundance in certain types of fish.

In one study, people with bipolar disorder -- previously known as manic depression -- had significantly fewer depressive episodes when their diets were supplemented with omega-3. And earlier research comparing 10 countries found that depression was much lower in areas where fish is a dietary staple.

GENERIC CONCEPT: RESPONSE

Passage No. 2.a; Prediction: relevant

Title: The IMF's response to the Asian crisis

The financial crisis that erupted in Asia in mid-1997 led to sharp declines in the currencies, stock markets, and other asset prices of a number of Asian countries; threatened these countries' financial systems; and disrupted their economies, with large contractions in activity that created a human crisis alongside the financial one. In pursuit of its immediate goal of restoring confidence in the region, the IMF took the following actions:

- helping the three countries most affected by the crisis-Indonesia, Korea, and Thailand-arrange programs of economic stabilisation and reform that could restore confidence and be supported by the IMF;
- approving in 1997 some SDR 26 billion or about US\$35 billion of IMF financial support for reform programs in Indonesia, Korea, and Thailand, and spearheading the mobilization of some US\$77 billion of additional financing from multilateral and bilateral sources in support of these reform programs. and

Appendix B – Generic Concept Analysis and Passage Selection

- intensifying its consultations with other members both within and outside the region that were affected by the crisis and needed to take policy steps to ward off the contagion effects, although not necessarily requiring IMF financial support.

Passage No. 2.b; Prediction: relevant
Title: HIV/AIDS: the epidemic and the national response
<p>China has not yet suffered a major HIV/AIDS epidemic, but the disease has established a foothold in the general population and, most experts agree, several factors now present are conducive to its spread.</p> <p>After the identification of China's first HIV cases, a National AIDS Committee was set up in October 1986, and this was followed in June 1987 by the establishment of a National Programme on HIV/AIDS Prevention and Control. In March 1990 a medium term prevention and control plan, in line with global policies modified to reflect local characteristics, was adopted by the Ministry of Health.</p> <p>... ..</p> <p>In October 1993 an external review of the AIDS control programme was conducted, to assess the relevance, adequacy and effectiveness of activities carried out under the medium term plan.</p> <p>A new, national strategic plan was formulated towards the end of 1994 to serve as a general framework up to the year 2000. This second national plan differs from the first in that management and prevention of other sexually transmitted diseases is included, and broader input is sought from sectors other than the health sector.</p> <p>.....</p>

Passage No. 2.c; Prediction: relevant
Title: Chinese response to terrorist strikes ranges from delight to disgust
<p>While American reactions to Tuesday's tragic events at the World Trade Centre and the Pentagon could be summed up as shock, grief and fury, Chinese perceptions have covered a broader spectrum, from joy to horror to no opinion at all. A group of poor young migrant</p>

Appendix B – Generic Concept Analysis and Passage Selection

workers pointed to the newspaper pictures of the Trade Centre's frame-by-frame demolition, laughing and remarking how "cool" it looked - as if it was Hollywood's latest special effect.

According to a teacher at Beijing Broadcasting Institute, one group of students held a celebration party, while others watched re-runs of the planes crash.

The perception that the incident was outrageous but justified pervaded among some. "While I'm completely shocked and disgusted, I also hope America starts to understand what this kind of loss of life feels like, to know how Iraq felt when their innocent civilians were bombed. I don't think this would have happened if American foreign policy weren't so aggressive," said one university graduate.

Passage No. 2.d 4; Prediction: irrelevant

Title: Lessons of the Asia Crisis

For more than a year, the global economic crisis, which began in Thailand July 1997 and spread rapidly throughout east Asia, and then to Russia and Latin America, has dominated the world economy. Almost every country in the world has been affected to some degree. In just a few short months, some went from robust growth to deep recession.

... ..

There is no single culprit for the problems that have beset the region. The economic situation in each country differed. But Global Economic Prospects concludes that the origins of the crisis lay fundamentally in the interaction between two things: the difficulties of domestic financial liberalisation and the problems associated with volatile inter-national capital markets.

Passage No. 2.e; Prediction: irrelevant

Title: Clinical management of HIV infection

Since it became available in the mid-1990s, highly active antiretroviral therapy (HAART) has significantly reduced the morbidity and mortality of HIV infection, and transformed the outlook for people living with the disease. It has revolutionised the management of HIV infection, turning it from being centred largely on the control of opportunistic infections and the provision of palliative care into a long-term strategy for controlling a chronic

Appendix B – Generic Concept Analysis and Passage Selection

condition.

Along with this considerable success, HAART has brought new challenges. The drug regimens are medically complex and can be difficult for patients to adhere to. The need to limit and manage adverse effects during long-term therapy, and to minimise the development of drug resistance and make the best use of available drugs, requires careful planning and constant attention.

Research into new antiretrovirals for the future is also important, and drugs with novel mechanisms of action are already being developed.

Passage No. 2.f; Prediction: irrelevant

Title: September 11th and the Bush administration

Clearly, one of the most critical questions of the twenty-first century concerns why the terrorist attacks of September 11, 2001 were not prevented. As I outline below, there are numerous aspects regarding the official stories about September 11th which do not fit with known facts, which contradict each other, which defy common sense, and which indicate a pattern of misinformation and coverup. The reports coming out of Washington do very little to alleviate these concerns.

For example, the Congressional report released on July 25, 2003 by a joint panel of House and Senate intelligence committees concluded that 9/11 resulted in C.I.A. and F.B.I. "lapses." While incompetence is frightening enough given a \$40 billion budget, it is simply not consistent with known facts. It is consistent with the reports from other government scandals such as the Iran Contra Affair which produced damage control and cover up but not answers to the more probing questions. But perhaps a comparison to Watergate is more apropos since we now have twenty-eight pages of this report, which the Bush Administration refuses to release. The report from the Federal Emergency Management Agency (FEMA) is believable unless you are seriously interested in the truth. Under more careful scientific scrutiny, it does not answer some very important questions.

... ..

GENERIC CONCEPT: REASON

Passage No. 3.a; Prediction: relevant

Appendix B – Generic Concept Analysis and Passage Selection

Title: Why should we drink so much water?

Water, after oxygen, is the second most important substance for human health. Water is a universal solvent and transport medium, and because of that it is the basis of all biological processes in the human body.

Water is mainly important for the digestive system, because it contributes to the constant supply and export of products and substances. The transport of nutrients can only take place through a solvent, and as such water acts as the main transport medium of nutrients. Water also attends heat regulation in our bodies. For humans it is of vital importance that the body temperature stays at a standard level. That is why we have to drink water, when we are infected with a fever. Water takes up heat and transports it out of the body while we are transpiring. We can survive without food for about 30 to 40 days, but we can only survive a few days without water. This is a factor that proves how important water is for us.

Passage No. 3.b; Prediction: relevant

Title: Why should we not drink seawater?

A long time ago a group of people were shipwrecked, but they had been able to save themselves and now they were floating around on a raft that they had constructed out of wooden boards from the ship that came floating up to the surface. They had no food and no water and they were so far away from the main land that they had no hope of being discovered any time soon. Because they were very thirsty, life became very hard on the raft after several days. Some of the passengers started drinking seawater in despair. But this made them die of dehydration pretty soon. How could this have happened while they were drinking water? It happened because seawater contains a lot of salt. When salt enters your body it will absorb a lot of water through a process called osmosis. This will cause the water content of your body to fall, which causes serious dehydration.

Passage No. 3.c; Prediction: relevant

Title: Origins of the Asian crisis

The financial crisis that erupted in Asia in mid-1997 led to sharp declines in the currencies,

Appendix B – Generic Concept Analysis and Passage Selection

stock markets, and other asset prices of a number of Asian countries; threatened these countries' financial systems; and disrupted their economies, with large contractions in activity that created a human crisis alongside the financial one.

The crisis unfolded against the backdrop of several decades of outstanding economic performance in Asia, and the difficulties that the East Asian countries face are not primarily the result of macroeconomic imbalances. Rather, they stemmed from weaknesses in financial systems and, to a lesser extent, governance. A combination of inadequate financial sector supervision, poor assessment and management of financial risk, and the maintenance of relatively fixed exchange rates led banks and corporations to borrow large amounts of international capital, much of it short-term, denominated in foreign currency, and unhedged. As time went on, this inflow of foreign capital tended to be used to finance poorer-quality investments.

Passage No. 3.d; Prediction: irrelevant

Title: The safe water system

The Centers for Disease Control and Prevention and the Pan American Health Organization/World Health Organization have developed a simple, inexpensive method of making water safe in the home. This method is called the Safe Water System. The Safe Water System has three parts:

- Water treatment with a locally-made, dilute bleach solution.
- Safe water storage in narrow-mouth containers with lids, and, if possible, a tap.
- Behavioral change techniques, including social marketing, community organization, and motivational interviewing to improve hygiene habits.

By the year 2015, the United Nations' Millennium Development Goal for Water aims to reduce by half the percentage of the world's population without safe water. If this goal is met, it will be a remarkable achievement, but hundreds of millions of people will still lack safe water.

Passage No. 3.e; Prediction: irrelevant

Title: Where does my drinking water come from?

Appendix B – Generic Concept Analysis and Passage Selection

Drinking water can come from either ground water sources (via wells) or surface water sources (such as rivers, lakes, and streams). Nationally, most water systems use a ground water source (80%), but most people (66%) are served by a water system that uses surface water. In addition, 10-20% of people have their own private well for drinking water. To find the source of your drinking water, check your annual water quality report or call your water supplier. You can get more information about specific watersheds by visiting EPA's Watershed Information Network. You can also learn more about EPA, state, and other efforts to protect sources of drinking water.

Passage No. 3.f; Prediction: irrelevant

Title: Lessons of the Asia crisis

For more than a year, the global economic crisis, which began in Thailand July 1997 and spread rapidly throughout east Asia, and then to Russia and Latin America, has dominated the world economy. Almost every country in the world has been affected to some degree. In just a few short months, some went from robust growth to deep recession.

The social consequences of this sharp downturn are already apparent: children dropping out of school, millions of people either falling back into poverty or coping with already desperate circumstances, and poorer health.

The crisis caught most economic forecasters off-guard. Even today, no one can predict how long the crisis will last or how deep it will be. But in the midst of this great uncertainty, it is important for us to have a sense of where the world economy is going, what has brought us to this juncture, and what we can do to improve our current outlook and to make another such global calamity less likely.

GENERIC CONCEPT: DEFINITION

Passage No. 4.a; Prediction: relevant

Title: What is cryptosporidium?

Cryptosporidium is a parasite that can cause watery diarrhoea and stomach ache, and possibly additional symptoms such as nausea, vomiting and fever, in humans who become infected. Healthy people who contract this infection almost always get better without any specific treatment and although symptoms can come and go for up to four weeks they usually subside in less. However, Cryptosporidium can cause severe illness in people with a

Appendix B – Generic Concept Analysis and Passage Selection

compromised immune system, such as those with HIV infection or taking drugs that cause immune suppression.

Passage No. 4.b; Prediction: relevant

Title: What is a psychiatrist?

A psychiatrist is a qualified medical doctor who has obtained additional qualifications to become a specialist in the diagnosis, treatment and prevention of mental illness and emotional problems. Because of their extensive medical and psychiatric training, psychiatrists are able to view illness in an integrated way by taking into consideration the related aspects of body and mind.

Psychiatrists are trained both to recognise and treat the effects of emotional disturbances on the body as a whole, as well as the effects of physical conditions on the mind. This is particularly important, as many emotional disturbances affect various parts of the body and physical illnesses can certainly affect the mind. A psychiatrist's medical and psychiatric training allows both the physical and emotional to be kept in perspective.

Passage No. 4.b; Prediction: relevant

Title: What is a psychiatrist?

A psychiatrist is a qualified medical doctor who has obtained additional qualifications to become a specialist in the diagnosis, treatment and prevention of mental illness and emotional problems. Because of their extensive medical and psychiatric training, psychiatrists are able to view illness in an integrated way by taking into consideration the related aspects of body and mind.

Psychiatrists are trained both to recognise and treat the effects of emotional disturbances on the body as a whole, as well as the effects of physical conditions on the mind. This is particularly important, as many emotional disturbances affect various parts of the body and physical illnesses can certainly affect the mind. A psychiatrist's medical and psychiatric training allows both the physical and emotional to be kept in perspective.

Passage No. 4.c; Prediction: relevant

Appendix B – Generic Concept Analysis and Passage Selection

Title: None
Liquid is a substance whose parts change their relative position on the slightest pressure, and therefore retain no definite form. Liquids have a fixed volume at any given pressure, but their shape is determined by the container in which it is contained. Liquids, in contrast to gases, cannot expand indefinitely to fill an expanding container, and are only slightly compressible by application of pressure.

Passage No. 4.d; Prediction: irrelevant
Title: None
Impaired communication and social interaction are the most fundamental symptoms of autism. Individuals with autism have serious problems interacting with others and often avoid eye contact.
<ul style="list-style-type: none">• As many as fifty percent of individuals with autism are non-verbal and up to eighty percent are intellectually challenged. A small percentage is gifted with extreme artistic or technical ability.• Common behaviours include seemingly purposeless repetitive behaviour, unusual responses to people or attachments to objects, resistance to change, and extreme sensory sensitivity.

Passage No. 4.e; Prediction: irrelevant
Title: None
Color theory is just that-a theory. However, there are many consistencies that arise from the blending of colored paint. What we've printed below owes a tremendous debt to Johannes Itten, the famous color theorist, fine artist, and former Bauhaus instructor. According to Itten "He who wants to become a master of color must see, feel, and experience each individual color in it's many endless combinations with all other colors".

Passage No. 4.f; Prediction: irrelevant
Title: None
Today, the use of penicillin and other antibiotics are common place. The various antibiotics

Appendix B – Generic Concept Analysis and Passage Selection

are used to treat a number of what are now common diseases and to prevent the onset of infections when our skin, our first barrier to fight off disease, is somehow broken through a simple cut or a more serious wound. It is something that we all take for granted, today. However, many diseases and simple wounds that are so easily treated today because of the availability of antibiotics has not always been available. Antibiotics are a relatively recent discovery and the first practical one, penicillin, was not available until the early 1940s.

GENERIC CONCEPT: EXPERIENCE

Passage No. 5.a; Prediction: relevant

Title: What is life like for an autistic adult?

For autistic adult, life is like hell on Earth, regardless of whatever positive testimonies you may read or hear. Even though you physically appear as an adult, you are emotionally still a child, and will still exhibit bizarre behavior for someone your age. It's not as if you're clinically insane and unable to reason, but you still feel trapped in a world of illusion and introspection. While everyone is outside enjoying a sunny day, you'll be inside your room researching the most insignificant minutiae on the internet, never moving from your seat except to go to the bathroom and eat.

The worst part about enduring autism as an adult is that you fail to meet basic "social norms." You don't "fit in." You never "get ahead." Your life becomes wasteful and unproductive. You may have dreams, but you'll never fulfill them because of the unusual, ineffable autistic reasons that "society" fails to understand. Don't expect to get married or have children, either, because you either won't care or you'll never find someone who shares the affection you may have for them.

Passage No. 5.b; Prediction: relevant

Title: Experience with Costa Rica

The temperature was mostly comfortable. We had only a few hot afternoons, and the nights were comfortable enough for sleeping without air conditioning. We needed only lightweight clothing for most days, a jacket for the nights in the mountains, and of course, rain gear. (The forests are rain forests, of course.)

It gets light about 6am, and dark about 6pm, plus or minus half an hour, year round. The

Appendix B – Generic Concept Analysis and Passage Selection

best wildlife is found just after dawn and just before and during dusk. Plan accordingly.

Although the towns seemed somewhat run down and poor, there was no one begging for money, no little kids trying to shine my shoes, and no aggressive street salesmen trying to sell me souvenirs. The people were wonderful to us.

About half of those we encountered spoke English. But with my very poor Spanish, sign language, drawing pictures, and writing out numbers, we managed just fine. We had reserved all lodging and some tours on-line in English in advance.

Passage No. 5.c; Prediction: relevant

Title: My Experience with ALS

It was a great shock to me to discover that I had motor neurone disease. I had never been very well co-ordinated physically as a child. I was not good at ball games, and my handwriting was the despair of my teachers. Maybe for this reason, I didn't care much for sport or physical activities. But things seemed to change when I went to Oxford, at the age of 17. I took up coxing and rowing. I was not Boat Race standard, but I got by at the level of inter-College competition.

In my third year at Oxford, however, I noticed that I seemed to be getting more clumsy, and I fell over once or twice for no apparent reason. But it was not until I was at Cambridge, in the following year, that my father noticed, and took me to the family doctor. He referred me to a specialist, and shortly after my 21st birthday, I went into hospital for tests. I was in for two weeks, during which I had a wide variety of tests. They took a muscle sample from my arm, stuck electrodes into me, and injected some radio opaque fluid into my spine, and watched it going up and down with x-rays, as they tilted the bed. After all that, they didn't tell me what I had, except that it was not multiple sclerosis, and that I was an a-typical case. I gathered however, that they expected it to continue to get worse, and that there was nothing they could do, except give me vitamins. I could see that they didn't expect them to have much effect. I didn't feel like asking for more details, because they were obviously bad.

... ..

Appendix B – Generic Concept Analysis and Passage Selection

Passage No. 5.d; Prediction: irrelevant
Title: Learning from past mistakes
<p>In 1991, the New Zealand Government introduced a points system for migration called the General Category. Later this became the General Skills Category.</p> <p>Anyone whose qualifications, age, etc, was "worth" sufficient points was granted permanent residence in New Zealand. With the advent of the points-based system, thousands of highly qualified people arrived. These new arrivals took up residence in their dream country, Godzone, and promptly found themselves unemployed.</p> <p>In the second half of 2004 Statistics New Zealand has released a 2001 census analysis of unemployment rates for degree-qualified new arrivals. For migrants from some countries, unemployment rates were very high.</p> <p>The statistics seem to show that unemployment rates fell with the length of time migrants spent in New Zealand. Unfortunately, migrant unemployment did not fall just because migrants were finding work.</p>

Passage No. 5.e; Prediction: irrelevant
Title: Costa Rica – Culture
<p>Costa Rica is noted more for its natural beauty and friendly people than for its culture. The overwhelming European influence erased almost all indigenous culture, and because Costa Rica was a country of subsistence agriculturalists until the middle of the 19th century, cultural activity has only begun to blossom in the last 100 years.</p> <p>By some estimates, more than 75% of Costa Ricans are Roman Catholics and 14% are evangelical Christians. In practice, most church attendance takes place at christenings, funerals and marriages. Blacks on the Caribbean coast tend to be Protestant, and there is a sprinkling of other denominations in San José, including a small Jewish community. Spanish is the official language, though English is understood in touristed areas. Many Caribbean blacks speak a lively dialect of English, known as Creole. Indigenous languages are spoken in isolated areas, primarily Bribri, which is estimated to be understood by about 10,000 people.</p>

Appendix B – Generic Concept Analysis and Passage Selection

Passage No. 5.f; Prediction: irrelevant
Title: Chinese cultural studies
Acknowledging the wisdom of Chinese proverbs, most anthologies of Chinese religion are organized by the logic of the three teachings (_sanjiao_) of Confucianism, Daoism, and Buddhism. Historical precedent and popular parlance attest to the importance of this threefold division for understanding Chinese culture. One of the earliest references to the trinitarian idea is attributed to Li Shiqian, a prominent scholar of the sixth century, who wrote that "Buddhism is the sun, Daoism the moon, and Confucianism the five planets." Li likens the three traditions to significant heavenly bodies, suggesting that although they remain separate, they also coexist as equally indispensable phenomena of the natural world. Other opinions stress the essential unity of the three religious systems. One popular proverb opens by listing the symbols that distinguish the religions from each other, but closes with the assertion that they are fundamentally the same:" The three teachings--the gold and cinnabar of Daoism, the relics of Buddhist figures, as well as the Confucian virtues of humanity and righteousness--are basically one tradition."

GENERIC CONCEPT: TREATMENT

Passage No. 6.a; Prediction: relevant
Title: History of the treatment of mental illness: a review
In many cases, doctors used drugs that not adequately tested, as well as electroconvulsive therapy (which was first called electroshock therapy), insulin shock therapy (which induced comas in patients by the injection of insulin), Metrazol (induced seizures), hydrotherapy (the wet sheet pack, the continuous bath), fever therapy and lobotomy.

Passage No. 6.b; Prediction: relevant
Title: Clinical treatments
Mood stabilising medications are helpful for people who have bipolar disorder (previously known as manic depression). Lithium carbonate can help reduce the recurrence of major depression and can help reduce the manic or 'high' episodes.

Passage No. 6.c; Prediction: relevant

Appendix B – Generic Concept Analysis and Passage Selection

Title: None
The mainstay of therapy is fluid and salt replacement. This may be accomplished through the use of oral rehydration salts or dilute Gatorade? in less severe cases, whereas IV fluids are often required in cases of severe dehydration. Antibiotics shorten the duration of diarrhea and thereby decrease fluid loss.

Passage No. 6.d; Prediction: irrelevant
Title: None
Description of Agent: Anthrax is a highly lethal infection caused by infection with the bacterium, <i>Bacillus anthracis</i> . In naturally-acquired cases, organisms usually gain entrance through skin wounds (causing a localised infection), but may be inhaled or ingested. Intentional release by belligerents or terrorist groups would presumably involve the aerosol route, as the spore form of the bacillus is quite stable and possess characteristics ideal for the generation of aerosols.

Passage No. 6.e; Prediction: irrelevant
Title: None
Diagnosis: Physical findings are non-specific in inhalational cases, with initial complaints of malaise, fever, headache, and possibly chest pain. A widened rib cage is sometimes seen on x-ray late in the course of illness, and correlates with a pathologic finding of hemorrhaging in the chest, the "classic" presentation of inhalational anthrax. The bacterium may be detected in a stain of blood and by blood culture late in the course of illness.

Passage No. 6.f; Prediction: irrelevant
Title: None
Arthritis can develop as a result of an infection. For example, bacteria that cause gonorrhea or Lyme disease can cause arthritis. Infectious arthritis can cause serious damage, but usually clears up completely with antibiotics. Scleroderma is a systemic disease that involves the skin, but may include problems with blood vessels, joints, and internal organs. Fibromyalgia syndrome is soft-tissue rheumatism that doesn't lead to joint

Appendix B – Generic Concept Analysis and Passage Selection

deformity, but affects an estimated 5 million Americans, mostly women. The approximate number of cases in the United States of some common forms of arthritis.

GENERIC CONCEPT: COMPOSITION

Passage No. 7.a; Prediction: relevant

Title: Overview of milk composition

Milk contains water, carbohydrate (lactose), fat, protein, minerals and vitamins. While each component can be discussed separately, it is important to remember that milk is secreted as a complex mixture of these components. The properties and importance of milk are greater and more complex than the sum of its individual component parts.

Consideration of individual milk components can be viewed from a number of perspectives, including:

- The biochemistry of each component
- The mechanisms of synthesis of each component
- The role of each component in defining the physicochemical properties of milk
- The function of each component in the mammary gland
- The importance of each component to the nursing young
- The importance of each component in milk and milk products as foods for humans

Passage No. 7.b; Prediction: relevant

Title: None

The human heart is primarily a shell. There are four cavities, or open spaces, inside the heart that fill with blood. Two of these cavities are called atria. The other two are called ventricles. The two atria form the curved top of the heart. The ventricles meet at the bottom of the heart to form a pointed base which points toward the left side of your chest. The left ventricle contracts most forcefully, so you can best feel your heart pumping on the left side of your chest.

Passage No. 7.c; Prediction: relevant

Title: None

1. DNA is made up of subunits which scientists called nucleotides.
2. Each nucleotide is made up of a sugar, a phosphate and a base.
3. There are 4 different bases in a DNA molecule:

Appendix B – Generic Concept Analysis and Passage Selection

<ul style="list-style-type: none">• adenine (a purine)• cytosine (a pyrimidine)• guanine (a purine)• thymine (a pyrimidine) <ol style="list-style-type: none">4. The number of purine bases equals the number of pyrimidine bases5. The number of adenine bases equals the number of thymine bases6. The number of guanine bases equals the number of cytosine bases <p>The basic structure of the DNA molecule is helical, with the bases being stacked on top of each other.</p>

Passage No. 7.d; Prediction: irrelevant
Title: Where— milk
Milk is not generally commercially available in the United States—and won't be until consumer demand increases. Raw milk can be purchased in some stores in California, Connecticut and New Mexico, but it is not necessarily from pasture-fed cows. However, in many states raw milk can be purchased at the farm and many concerned consumers are forming cooperatives designed to support conscientious dairy farmers and obtain milk directly from the farm. The solution to restrictive state laws is a cow-share program in which farmers keep and milk cows owned by individuals. These are being set up in many states.

Passage No. 7.e; Prediction: irrelevant
Title: Heart disease
Cardiovascular disease (CVD), including heart disease and stroke, remains the leading cause of death in the United States despite improvements in prevention, detection, and treatment. CVD is no longer thought of as a disease that primarily affects men as they age. It is a killer of people in the prime of life, with more than half of all deaths occurring among women. Cardiovascular diseases remain the leading cause of disability among working adults. Stroke alone accounts for the disability of more than a million Americans. The economic impact on the health system grows larger as the population ages. In 2001, the estimated cost of health care expenditures and lost productivity attributable to cardiovascular

Appendix B – Generic Concept Analysis and Passage Selection

diseases was \$298 billion.

Passage No. 7.f; Prediction: irrelevant

Title: None

If we heat up a tube of DNA dissolved in water, the energy of the heat can pull the two strands of DNA apart (there's a critical temperature called the T_m at which this happens). This process is called 'denaturation'; when we've 'denatured' the DNA, we have heated it to separate the strands.

The two strands still have the same nucleotide sequences, however, so they are still complementary. If we cool the tube again, then in the course of the normal, random molecular motion they'll eventually bump into each other ... and stick tightly, reforming double-stranded DNA. This process is called 'annealing' or 'hybridization', and it is very specific; only complementary strands will come together if it is done right. This process is used in many crime labs to identify specific strands of DNA in a mixture.

B.2 A SAMPLE QUESTIONNAIRE

In this experiment, we give you 7 pairs of short passages, one pair per page. For each pair of passages, we give a concept displayed at the top of the page, and you are required to choose whichever passage you believe is more *relevant* to the given concept. For example, if the given concept is 'cause', then a passage about the cause (or causes) of an event would be relevant.

You must choose ONE and ONLY ONE passage from each pair and *tick the box* below the passage that you choose. It is important that when they are BOTH relevant, or BOTH irrelevant, you should still tick the one that is MORE relevant.

Please go through the passages quickly and make your choice, the experiment will take about 15 minutes.

Appendix B – Generic Concept Analysis and Passage Selection

CAUSE

Passage 1

Some suicides are the result of impulsive decisions based on a situation that seems hopeless - loss of a job, divorce, or a breakup with one's girlfriend or boyfriend. Suicide attempts triggered by major disappointments, such as romantic rejection, problems with peers, or failing a big exam, are common among depressed teenagers, who haven't had the life experience to realize that these "injuries" heal with time.

Ninety percent of the people who commit suicide have a mental or substance abuse disorder (or both). More than half of the people who kill themselves are seriously or clinically depressed.

□

Passage 2

No-one knows all the reasons why people become mentally ill. Some people have a 'chemical imbalance' which affects how their brain works. This makes them have strange thoughts or feelings, or behave oddly. They may need to take medication to help their brain work better. For other people, something might happen in their life which is very stressful, such as a death of someone very close, and this may trigger a mental illness.

Mental illness doesn't normally start out of the blue. It usually develops slowly. But some people do get a mental illness suddenly, such as when someone has a psychotic illness.

□

Appendix B – Generic Concept Analysis and Passage Selection

TREATMENT

Passage 1

Anthrax is a highly lethal infection caused by infection with the bacterium, *Bacillus anthracis*. In naturally-acquired cases, organisms usually gain entrance through skin wounds (causing a localised infection), but may be inhaled or ingested. Intentional release by belligerents or terrorist groups would presumably involve the aerosol route, as the spore form of the bacillus is quite stable and possess characteristics ideal for the generation of aerosols.

□

Passage 2

Arthritis can develop as a result of an infection. For example, bacteria that cause gonorrhoea or Lyme disease can cause arthritis. Infectious arthritis can cause serious damage, but usually clears up completely with antibiotics. Scleroderma is a systemic disease that involves the skin, but may include problems with blood vessels, joints, and internal organs. Fibromyalgia syndrome is soft-tissue rheumatism that doesn't lead to joint deformity, but affects an estimated 5 million Americans, mostly women. The approximate number of cases in the United States of some common forms of arthritis.

□

COMPOSITION

Passage 1

Cardiovascular disease (CVD), including heart disease and stroke, remains the leading cause of death in the United States despite improvements in prevention, detection, and treatment. CVD is no longer thought of as a disease that primarily affects men as they age. It is a killer of people in the prime of life, with more than half of all deaths occurring among women.

Cardiovascular diseases remain the leading cause of disability among working adults. Stroke alone accounts for the disability of more than a million Americans. The economic impact on the health system grows larger as the population ages. In 2001, the estimated cost of health care expenditures and lost productivity attributable to cardiovascular diseases was \$298 billion.

□

Passage 2

1. DNA is made up of subunits which scientists called nucleotides.
2. Each nucleotide is made up of a sugar, a phosphate and a base.
3. There are 4 different bases in a DNA molecule:
 - adenine (a purine)
 - cytosine (a pyrimidine)
 - guanine (a purine)
 - thymine (a pyrimidine)
4. The number of purine bases equals the number of pyrimidine bases
5. The number of adenine bases equals the number of thymine bases
6. The number of guanine bases equals the number of cytosine bases

The basic structure of the DNA molecule is helical, with the bases being stacked on top of each other.

□

RESPONSE

Passage 1

The financial crisis that erupted in Asia in mid-1997 led to sharp declines in the currencies, stock markets, and other asset prices of a number of Asian countries; threatened these countries' financial systems; and disrupted their economies, with large contractions in activity that created a human crisis alongside the financial one. In pursuit of its immediate goal of restoring confidence in the region, the IMF took the following actions:

- helping the three countries most affected by the crisis-Indonesia, Korea, and Thailand-arrange programs of economic stabilization and reform that could restore confidence and be supported by the IMF;
- approving in 1997 some SDR 26 billion or about US\$35 billion of IMF financial support for reform programs in Indonesia, Korea, and Thailand, and spearheading the mobilization of some US\$77 billion of additional financing from multilateral and bilateral sources in support of these reform programs. and
- intensifying its consultations with other members both within and outside the region that were affected by the crisis and needed to take policy steps to ward off the contagion effects, although not necessarily requiring IMF financial support.

□

Passage 2

Clearly, one of the most critical questions of the twenty-first century concerns why the terrorist attacks of September 11, 2001 were not prevented. As I outline below, there are numerous aspects regarding the official stories about September 11th which do not fit with known facts, which contradict each other, which defy common sense, and which indicate a pattern of misinformation and coverup. The reports coming out of Washington do very little to alleviate these concerns.

Appendix B – Generic Concept Analysis and Passage Selection

For example, the Congressional report released on July 25, 2003 by a joint panel of House and Senate intelligence committees concluded that 9/11 resulted in C.I.A. and F.B.I. "lapses." While incompetence is frightening enough given a \$40 billion budget, it is simply not consistent with known facts. It is consistent with the reports from other government scandals such as the Iran Contra Affair which produced damage control and cover up but not answers to the more probing questions. But perhaps a comparison to Watergate is more apropos since we now have twenty-eight pages of this report, which the Bush Administration refuses to release. The report from the Federal Emergency Management Agency (FEMA) is believable unless you are seriously interested in the truth. Under more careful scientific scrutiny, it does not answer some very important questions.

.....

□

Appendix B – Generic Concept Analysis and Passage Selection

REASON

Passage 1

A long time ago a group of people were shipwrecked, but they had been able to save themselves and now they were floating around on a raft that they had constructed out of wooden boards from the ship that came floating up to the surface. They had no food and no water and they were so far away from the main land that they had no hope of being discovered any time soon.

Because they were very thirsty, life became very hard on the raft after several days. Some of the passengers started drinking seawater in despair. But this made them die of dehydration pretty soon. How could this have happened while they were drinking water? It happened because seawater contains a lot of salt. When salt enters your body it will absorb a lot of water through a process called osmosis. This will cause the water content of your body to fall, which causes serious dehydration.

□

Passage 2

For more than a year, the global economic crisis, which began in Thailand July 1997 and spread rapidly throughout east Asia, and then to Russia and Latin America, has dominated the world economy. Almost every country in the world has been affected to some degree. In just a few short months, some went from robust growth to deep recession.

The social consequences of this sharp downturn are already apparent: children dropping out of school, millions of people either falling back into poverty or coping with already desperate circumstances, and poorer health.

The crisis caught most economic forecasters off-guard. Even today, no one can predict how long the crisis will last or how deep it will be. But in the midst of this great uncertainty, it is important for us to have a sense of where the world economy is going, what has brought us to this juncture, and what we can do to improve our current outlook and to make another such global calamity less likely.

□

Appendix B – Generic Concept Analysis and Passage Selection

TREATMENT

Passage 1

Mood stabilising medications are helpful for people who have bipolar disorder (previously known as manic depression). Lithium carbonate can help reduce the recurrence of major depression and can help reduce the manic or 'high' episodes.

□

Passage 2

Anthrax is a highly lethal infection caused by infection with the bacterium, *Bacillus anthracis*. In naturally-acquired cases, organisms usually gain entrance through skin wounds (causing a localized infection), but may be inhaled or ingested. Intentional release by belligerents or terrorist groups would presumably involve the aerosol route, as the spore form of the bacillus is quite stable and possess characteristics ideal for the generation of aerosols.

□

Appendix B – Generic Concept Analysis and Passage Selection

RESPONSE

Passage 1

For more than a year, the global economic crisis, which began in Thailand July 1997 and spread rapidly throughout east Asia, and then to Russia and Latin America, has dominated the world economy. Almost every country in the world has been affected to some degree. In just a few short months, some went from robust growth to deep recession.

.....

There is no single culprit for the problems that have beset the region. The economic situation in each country differed. But Global Economic Prospects concludes that the origins of the crisis lay fundamentally in the interaction between two things: the difficulties of domestic financial liberalisation and the problems associated with volatile inter-national capital markets.

□

Passage 2

Since it became available in the mid-1990s, highly active antiretroviral therapy (HAART) has significantly reduced the morbidity and mortality of HIV infection, and transformed the outlook for people living with the disease. It has revolutionised the management of HIV infection, turning it from being centred largely on the control of opportunistic infections and the provision of palliative care into a long-term strategy for controlling a chronic condition.

Along with this considerable success, HAART has brought new challenges. The drug regimens are medically complex and can be difficult for patients to adhere to. The need to limit and manage adverse effects during long-term therapy, and to minimise the development of drug resistance and make the best use of available drugs, requires careful planning and constant attention.

Research into new antiretrovirals for the future is also important, and drugs with novel mechanisms of action are already being developed.

□

Appendix B – Generic Concept Analysis and Passage Selection

Thank you very much for your help!

Please also answer the following questions.

Are you a native English Speaker? Yes No

How long did it take for you to finish the experiment?

APPENDIX C

STUDY THE STRUCTURE OF EXTENDED TOPIC

This appendix contains materials used in the experiment described in section 4.4. The purpose of the experiment is to verify the structure of extended topic; in particular, it aims to test whether the concepts used in generic topics are generally more general than the concepts used in specific topics.

C.1 SAMPLE TOPIC EXPRESSIONS

Below are a set of sample topic expressions used in the experiment, organised into four categories, including 'CL+Describe', 'CL+Present', 'Physics+Describe' and 'Physics+Present'. The example topic expressions in category 'CL+Describe' are collected by searching cue phrases such as 'this paper describes' and 'we describe' in a collection of academic papers in Computational Linguistics. Topic expressions in other three categories are collected in similar way with Computational Linguistics Papers being replaced by Physics Papers or with 'Describe' being replaced by 'Present'.

'CL + DESCRIBE'

[1] This paper describes a novel computer-aided procedure for generating multiple-choice tests from electronic instructional documents.

[2] This paper describes a distributional approach to the semantics of verb-particle constructions.

[3] This paper describes log-linear parsing models for Combinatory Categorical Grammar (CCG).

[4] This paper describes preliminary work in exploring the relative effectiveness of speech versus text based tutorial dialogue systems.

[5] This paper describes classification of typed student utterances within AutoTutor, an intelligent tutoring system.

[6] This Section 4 describes the method used to produce the TAGs that are the basis of our experiments.

[7] This Section 4 describes the classifier algorithm.

[8] This section describes a counting algorithm based on general weighted automata algorithms.

Appendix C – Study the Structure of Extended Topic

[9] This section describes a simple and efficient method for constructing class-based language models where each class may represent an arbitrary (weighted) regular language.

[10] This Section 3 describes ways of translation with a grammatical relation dictionary and k-nearest neighbor learning method.

‘CL+ PRESENT’

[1] Section 6 presents evaluation of the classifier on AutoTutor sessions.

[2] Section 2 presents AutoTutor.

[3] Section 3 presents two methods of representing the feature space.

[4] This paper presents a new method for estimating automatically the stability classes that indicate index of words popularity with time-series variation based on frequency change in past texts data.

[5] Abney also presents a greedy algorithm that maximises agreement on unlabelled data.

[6] Section 2 presents the ontology-based framework for linguistic annotation.

[7] Section 4 presents CREAM.

[8] Section 3 presents an overview of our work and distinguishes it from previous work.

[9] This paper presents a primarily data-driven Chinese word segmentation system and its performances on the closed track using two corpora at the first international Chinese word segmentation bakeoff.

[10] This paper presents a Named Entity Extraction (NEE) system for the CoNLL-2003 shared task competition.

‘PHYSICS + DESCRIBE’

[1] This paper describes a novel technique of local surface fabrication using a scanning probe microscope with a thermally pulled micropipette probe.

[2] This paper describes a novel wafer bonding technique using microwave heating of parylene intermediate layers.

[3] This paper describes a micropump composed of a piezoelectric PZT unimorph and one-way parylene valves.

[4] This paper describes a micromachined magnetic field sensor based on magnetic resonant structures.

Appendix C – Study the Structure of Extended Topic

[5] This paper describes both the more than 30-year-old history and the present state of development and applications.

[6] This paper describes spray characteristics of a low-pressure common rail injector which is intended for use in an HCCI engine.

[7] This paper describes a two-factor model for a diversified market index using the growth optimal portfolio with a stochastic and possibly correlated intrinsic timescale.

[8] This paper describes a method for calibrating the angles of transparent prism-type optical elements.

[9] This paper describes a new method for fabricating a gas sensor composed of multi-wall carbon nanotubes (MWCNTs) using dielectrophoresis (DEP).

[10] This paper describes further developments of a near-wall second moment turbulence closure of Manceau and Hanjalić, and its application to the imposed system rotation around three orthogonal and non-orthogonal rotating axes on turbulent channel flow.

‘PHYSICS+PRESENT’

[1] This paper presents a novel fabrication process for a tapered hollow metallic microneedle array using backside exposure of SU-8, and analytic solutions of critical buckling of a tapered hollow microneedle.

[2] This paper presents the first micropumps assembled using polymeric lamination technology.

[3] This paper presents a novel pattern transfer process of LIGA and UV-LIGA MEMS onto CMOS chips using polydimethylsiloxane (PDMS) replication and electroplating-based post-IC integration techniques.

[4] This paper presents detailed calculations of the flow.

[5] This paper presents a new process methodology developed for this purpose, which consists of varying the projection aperture during machining by applying relative motion to two overlapping masks.

[6] This paper presents the adaptation of a conventional injection molding process to the mass replication of polymeric microstructures with appropriate mold design and process control.

[7] This paper presents numerical and experimental results of active compensation of thermal deformation of a composite beam using piezoelectric ceramic actuators.

[8] This paper presents a mathematical model to design and fabricate micro-ball lens array using thermal reflow in two polymer layers.

[9] This paper presents a microcomputer-based ultrasonic temperature sensor system to measure the temperature of an air conditioner (AC) in an automobile.

[10] This paper presents a touch probe system with improved sensitivity and repeatability.

C.2 ALGORITHM FOR EXTRACTING HEAD NOUNS FROM A NOUN PHRASE

Function: This algorithm automatically extracts the head nouns and the non-head nouns from a noun phrase based on the word part of speech (POS) tags and put them into two lists.

Input: The input of this algorithm is a sentence and a pointer marking the starting point of the noun phrase in the sentence. The sentence is annotated with some syntactic information generated by NLPProcessor. Below is the syntactic annotation of sentence ‘this paper presents detailed calculations of the flow’, where the C attribute of each word encodes its POS. Other attributes of a word is not important as we only apply the POS tag.

```
<S><W chunk='NGstart' C='DT' L='SL' T='w' S='Y'>This</W> <W chunk='NGin'
C='NN'>paper</W> <W chunk='NGend' C='NNS'>presents</W> <W chunk='VGstart_end'
C='VBD'>detailed</W> <W chunk='NGstart_end' C='NNS'>calculations</W> <W C='IN'>of</W>
<W chunk='NGstart' C='DT'>the</W> <W chunk='NGend' C='NN'>flow</W><W C='.'
T='.'>.</W></S>
```

Algorithm:

```
private Vector headnounList;

//A global variable which keeps all the extracted head nouns

private Vector nonHeadnounList;

//A global variable which keeps all the extracted non-head nouns

private void topicanalysis (NodeList wordlist, int position){

/*
```

WordList is a data structure that keeps the words and their syntactic information of a topic expression in the original order.

Position encodes the starting point of the noun phrase in the topic expression.

Appendix C – Study the Structure of Extended Topic

*/

```
boolean findFirst=true;

//signal a status in which the first head noun is being looked for.

boolean findRest=false;

/* signal a status in which a head noun is still being looked for after the first head
noun has been spotted. This status is often triggered by the identification of a
parallel structure marker such as word 'and'. */

String keepKeyWordValue=null;

//This variable stores the most recently examined noun during the iteration

int headnounPosition=0;

//This variable keeps the position of the last spotted head noun

Stemmer stem=new Stemmer();

//A simple algorithm that changes the plural form of a noun to the normal form

for (int ii=position+1;ii<wordList.getLength();ii++){

/* starting from the first word of the noun phrase, examine words one by one. ii
marks the position of the word under examination in current iteration.*/

    //get current word under examination

    Node keyWord=wordList.item(ii).getFirstChild();

    String keyWordValue=keyWord.getNodeValue();

    //get the POS information

    NamedNodeMap attributes=wordList.item(ii).getAttributes();

    Node posNode=attributes.getNamedItem("C");

    String pos=posNode.getNodeValue();

    //stem the word

    {

        stem.add(keyWordValue.toCharArray(),keyWordValue.length());

        stem.stem();

    }

}
```

Appendix C – Study the Structure of Extended Topic

```
        keyWordValue=stem.toString();
    }
    if(judgeWordPos(pos, "POS_NOUN")){
//Check whether the current word is a noun
        if(findFirst| |findRest){
            if(keepKeyWordValue!=null){
                nonHeadnounList.addElement(keepKeyWordValue);
                /*the noun kept by variable keepKeyWordValue is
                considered as a modifier before the head noun. */
            }
            //store information of the current noun
            keepKeyWordValue=keyWordValue;
            headWordPosition=ii;
        }else{
            nonHeadnounList.addElement(keyWordValue);
        }
    }else{//the current word is not a Noun
        if((findFirst| |findRest)&&(keepKeyWordValue!=null)){
            /* This indicates that the process have reached the end of the first
            noun group, and the last noun of the first noun group is considered
            as a head noun */
            headnounList.addElement(keepKeyWordValue);
            if(findFirst) findFirst=false;
            if(findRest) findRest=false;
            keepKeyWordValue=null;
        }
    }
}
```


Appendix C – Study the Structure of Extended Topic

[1] a novel fabrication process for a tapered hollow metallic microneedle array using backside exposure of SU-8, and analytic solutions of critical buckling of a tapered hollow microneedle

Head Noun: process solutions

Head Noun (extracted): process

[2] the first micropumps assembled using polymeric lamination technology

Head Noun: micropumps

Head Noun (extracted): micropumps

[3] a novel pattern transfer process of LIGA and UV-LIGA MEMS onto CMOS chips using polydimethylsiloxane (PDMS) replication and electroplating-based post-IC integration techniques.

Head Noun: process

Head Noun (extracted): process

[4] detailed calculations of the flow.

Head Noun: calculations

Head Noun (extracted): calculations

[5] a new process methodology developed for this purpose, which consists of varying the projection aperture during machining by applying relative motion to two overlapping masks.

Head Noun: methodology

Head Noun (extracted): methodology

[6] the adaptation of a conventional injection molding process to the mass replication of polymeric microstructures with appropriate mold design and process control.

Head Noun: adaptation

Head Noun (extracted): adaptation

[7] numerical and experimental results of active compensation of thermal deformation of a composite beam using piezoelectric ceramic actuators.

Head Noun: results

Head Noun (extracted): results

Appendix C – Study the Structure of Extended Topic

[8] a mathematical model to design and fabricate micro-ball lens array using thermal reflow in two polymer layers.

Head Noun: model

Head Noun (extracted): model

[9] a microcomputer-based ultrasonic temperature sensor system to measure the temperature of an air conditioner (AC) in an automobile.

Head Noun: system

Head Noun (extracted): system

[10] a touch probe system with improved sensitivity and repeatability.

Head Noun: system

Head Noun (extracted): system

[11] This paper presents both experimental and theoretical studies on the atomic structure changes of monocrystalline silicon brought about by surface nano-modification.

Head Noun: Studies

Head Noun (extracted): Studies

[12] the novel concept of a silicon-based coupling platform integrated with an electrostatic-driven out-of-plane optical switch, a self-parking framework, a flat-topped mesa, a v-groove and a microball lens array.

Head Noun: concept

Head Noun (extracted): concept

[13] a technique for measuring and quantifying the dielectrophoretic collection of sub-micron particles on planar microelectrode arrays.

Head Noun: technique

Head Noun (extracted): technique

[14] a new algorithm to calibrate the option pricing model.

Head Noun: algorithm

Head Noun (extracted): algorithm

Appendix C – Study the Structure of Extended Topic

[15] the different steps involved in designing, building and testing an intelligent damper, which is originally a classic passive damper retrofitted with electro-rheological (ER) technology, that can be used for semi-active car suspensions.

Head Noun: steps

Head Noun (extracted): steps

[16] two novel tracer gas based techniques that have been developed to measure the axial variation of wall flow through the porous channel walls of a diesel particulate filter.

Head Noun: technique

Head Noun (extracted): gas

[17] the design and experimental results of control of an SMA actuator using pulse width modulation (PWM) to reduce the energy consumption by the SMA actuator.

Head Noun: Design, results

Head Noun (extracted): Design, results

[18] a brief history of the technique and its application to archaeology, describes the physics behind the analytical method, and explains how the method is generally employed to determine the sources of archaeological materials.

Head Noun: history, application

Head Noun (extracted): history

[19] a novel technique for the fabrication of a micro humidity sensor with suspending structures.

Head Noun: technique

Head Noun (extracted): technique

[20] the development of a reduced-order macro-model for the double-gimballed electrostatic torsional micromirror using the hierarchical circuit-based approach.

Head Noun: development

Head Noun (extracted): development

[21] a new microlens array fabrication method that controls the printing gap in the UV lithography process.

Appendix C – Study the Structure of Extended Topic

Head Noun: method

Head Noun (extracted): method

[22] techniques and methods used in the C2RMF laboratory for manganese oxide pigments.

Head Noun: techniques, methods

Head Noun (extracted): techniques, methods

[23] a micromachined in-plane tunable optical filter using the thermo-optic effect of crystalline silicon.

Head Noun: filter

Head Noun (extracted): filter

[24] an overview of the different methods of evaluating these transmission data, leading to values for the complex refractive index.

Head Noun: overview

Head Noun (extracted): overview

[25] a numerical investigation of the thermal behavior of an artificial anal sphincter using shape memory alloys (SMAs) proposed by the authors.

Head Noun: investigation

Head Noun (extracted): investigation

[26] a study of an arrangement consisting of two matched pairs of piezoelectric transducers bonded on a cantilever beam for vibration control.

Head Noun: study

Head Noun (extracted): study

[27] a short review of the wider state-of-the-art development of these single cold trapped ion frequency standards.

Head Noun: review

Head Noun (extracted): review

[28] a simple model of the fragment in the cathode electrical arc root taking into account the physical phenomena occurring on the cathode surface and the sheath.

Appendix C – Study the Structure of Extended Topic

Head Noun: model

Head Noun (extracted): model

[29] the principle and design of a prototype, digital imaging based, instrumentation system that can measure size distribution of particles that are largely opaque.

Head Noun: principle, design

Head Noun (extracted): principle, design

[30] the results of 2D-finite-element simulations of the periodic change in capacitance between two periodic structures.

Head Noun: results

Head Noun (extracted): results

[31] the application of digital imaging and image processing techniques for the quantitative characterization of diesel sprays.

Head Noun: application

Head Noun (extracted): application

[32] a possible explanation of this observation based on an analysis of strain profiles and resistivity behaviour difference in resistors with different thicknesses subjected to temperature variation.

Head Noun: explanation

Head Noun (extracted): explanation

[33] an experimental investigation and theoretical modelling of shape formation of high aspect ratio columns and lines fabricated by LECD.

Head Noun: investigation, modelling

Head Noun (extracted): investigation, modelling

[34] a device that integrates a micromachined flow cytometer with two embedded etched optic fibers in order to carry out on-line detection of particles and cells.

Head Noun: device

Head Noun (extracted): device

[35] data that may aid physical interpretation of leakage currents on insulators suffering pollution under physical conditions.

Head Noun: data

Head Noun (extracted): data

C.4 MOST FREQUENT TERMS IN HEAD NOUN AND NON-HEAD NOUN LISTS

Term	Frequency
Method	37
System	26
Approach	26
Algorithm	21
Model	20
Experiment	17
Work	16
Result	12
Task	10

Term	Frequency
Feature	8
Set	7
Design	7
Pattern	6
Detail	6
Corpus	6
Procedure	5
Name	5

TABLE C-1. TOP 10 PERCENT HEAD NOUN TERMS AND FREQUENCIES IN 'CL+DESCRIBE'

Term	Frequency
System	50
Method	31
Translation	26
Text	23
Task	22
Word	22
Feature	21
Information	21
Language	21
Algorithm	19
Model	18
Data	18
Evaluation	17
Dialogue	15
Experiment	15

Term	Frequency
Corpus	14
Detail	13
Section	12
Corpora	12
Approach	12
Speech	12
Technique	12
Verb	12
Result	12
Retrieval	11
Learning	11
Structure	11
Student	11
Dependency	10

TABLE C-2. TOP 5 PERCENT NON-HEAD NOUN TERMS AND FREQUENCIES IN 'CL+DESCRIBE'

Term	Frequency
Approach	53
Method	45
Result	41
Work	25

Term	Frequency
Model	16
Conclusion	11
Algorithm	11
Evaluation	10

Appendix C – Study the Structure of Extended Topic

Term	Frequency
System	9
Section	7
Framework	7

Term	Frequency
Experiment	7
Technique	7

TABLE C-3. TOP 10 PERCENT HEAD NOUN TERMS AND FREQUENCIES IN 'CL+PRESENT'

Term	Frequency
Word	38
Corpus	25
System	21
Algorithm	18
Lexicon	17
Present	16
Learning	16
Section	15
Text	15
Data	14

Term	Frequency
Evaluation	14
Corpora	13
Information	13
Result	13
Task	13
Class	13
Classification	12
Constraint	12
Model	12

TABLE C-4. TOP 5 PERCENT NON-HEAD NOUN TERMS AND FREQUENCIES IN 'CL+PRESENT'

Term	Frequency
Method	20
Development	13
Design	12
Technique	7

Term	Frequency
System	6
Application	5
Operation	4

TABLE C-5. TOP 10 PERCENT HEAD NOUN TERMS AND FREQUENCIES IN 'PHYSICS+DESCRIBE'

Term	Frequency
System	14
Structure	13
Flow	13
Measurement	12
Sensor	9
Analysis	9
Probe	9
Technique	7
Silicon	7
Data	6

Term	Frequency
Wave	6
Velocity	6
Fabrication	6
Laser	5
Process	5
Device	5
Substrate	5
Acquisition	4
Use	4
Field	4

TABLE C-6. TOP 5 PERCENT NON-HEAD NOUN TERMS AND FREQUENCIES IN 'PHYSICS+DESCRIBE'

Term	Frequency
Result	21
Study	20
Method	18
Design	16
Model	14

Term	Frequency
Analysis	12
System	9
Approach	8
Investigation	7
Technique	7

TABLE C-7. TOP 10 PERCENT HEAD NOUN TERMS AND FREQUENCIES IN 'PHYSICS+PRESENT'

Appendix C – Study the Structure of Extended Topic

Term	Frequency
Control	26
Actuator	21
System	16
Measurement	15
Structure	14
Technique	13
Temperature	13
Wave	11
Method	11
Array	10
Process	10
Shape	10
Design	10
Field	10
Sensor	10

Term	Frequency
Vibration	9
Material	9
Surface	9
Particle	9
Size	8
Beam	8
Effect	8
Flow	8
Application	8
Plate	7
Memory	7
Imaging	6
Force	6
Type	6
Use	6

TABLE C-8. TOP 5 PERCENT NON-HEAD NOUN TERMS AND FREQUENCY IN 'PHYSICS+PRESENT'

C.5 QUESTIONNAIRE FOR PROBING TERM GENERALITY

The following words are collected from academic papers in Applied Physics and Computational Linguistics. For each word, please make the choice between: physics term, linguistic term and scientific term in general.

- A physics term would refer to **some special objects of study**, or typical method, tool in Applied Physics or related areas (general physics, Electronic Engineering, etc).
- A linguistic term would refer to **some special objects of study**, or typical method, algorithm in Computational Linguistics or related areas (Linguistics, Computer Science, etc).
- A scientific term is applied to scientific research in general.

Please give a quick judgement according to your intuition. For every single word, please tick only ONE box.

Key Word	Physics	Linguistics	Scientific Term
Development	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Design	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Dialogue	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Technique	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
System	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Appendix C – Study the Structure of Extended Topic

Set	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Corpus	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Name	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Architecture	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Motivation	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Translation	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Text	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Word	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Flow	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Information	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Language	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Data	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Evaluation	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Section	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Corpora	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Method	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Conclusion	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Example	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Problem	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Idea	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Lexicon	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Present	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Fabrication	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Learning	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Class	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Classification	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Constraint	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Appendix C – Study the Structure of Extended Topic

Dependency	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Result	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Device	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Application	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Approach	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Experiment	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Model	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
State	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Overview	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Process	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
set-up	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Procedure	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Theory	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Structure	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Characteristic	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Measurement	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Analysis	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Probe	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Data	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Wave	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Laser	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Substrate	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Acquisition	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Use	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Field	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Study	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Investigation	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Appendix C – Study the Structure of Extended Topic

Methodology	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Solution	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Silicon	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Algorithm	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Control	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Simulation	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Principle	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Velocity	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Mechanism	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Proof	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Parser	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Sensor	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Classifier	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Framework	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Speech	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Verb	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Learning	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Student	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Dependency	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Question	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Platform	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Training	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Formalism	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Construction	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Chart	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Number	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
View	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Appendix C – Study the Structure of Extended Topic

Recognition	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Extraction	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Particle	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Answer	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Feature	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Location	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Case	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Segmentation	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Module	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Actuator	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Temperature	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Array	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Shape	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Vibration	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Retrieval	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Material	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Surface	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Size	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

C.6 INTER-RATER AGREEMENT TEST

The Kappa coefficient is an index for calculating inter-rater agreement. The formula is as follows.

$$\hat{k} = \frac{P_A - P_E}{1 - P_E} \tag{C.1}$$

Where P_A denotes the percentage of agreement, it is calculated by formula C.2 if there are two annotators. In formula C.2, i denotes a category, and P_{ii} means the percentage of cases which are assigned to the same category by two annotators. P_E refers to the possibility of agreement by chance. There are different versions of formulas for calculating the value P_E .

Appendix C – Study the Structure of Extended Topic

Formula C.3 and formula C.4 are Cohen’s (1960) version and the Fleiss’ (1971) version respectively. In these two formulas, P_{i_1} means the percentage of cases being assigned to a category i by annotator 1, and P_{i_2} means the percentage of terms being assigned to a category i by annotator 2.

$$P_A = \sum_i P_{ii} \tag{C.2}$$

$$P_{E-COHEN} = \sum_i P_{i_1} * P_{i_2} \tag{C.3}$$

$$P_{E-FLEISS} = \sum_i \left(\frac{P_{i_1} + P_{i_2}}{2} \right)^2 \tag{C.4}$$

The following tables show the result of our term generality probing experiment. The rows represent the judgement of one subject and the columns represent another.

(1) Subject 1 vs. Subject 2

	<i>Physics Term</i>	<i>Linguistic Term</i>	<i>General Scientific Term</i>	<i>Total</i>
<i>Physics Term</i>	9	0	11	20
<i>Linguistic Term</i>	0	19	1	20
<i>General Scientific Term</i>	4	21	38	63
<i>Total</i>	13	40	50	103

$$P_A = 0.6407$$

$$P_{E-COHEN} = 0.3968$$

$$\hat{k} = 40.4\%$$

$$P_{E-FLEISS} = 0.4114$$

$$\hat{k} = 38.95\%$$

(2) Subject 1 vs. Subject 3

	<i>Physics Term</i>	<i>Linguistic Term</i>	<i>General Scientific Term</i>	<i>Total</i>
<i>Physics Term</i>	11	0	1	12
<i>Linguistic Term</i>	1	14	13	28
<i>General Scientific Term</i>	7	6	50	63
<i>Total</i>	19	20	64	103

$$P_A = 0.7282$$

Appendix C – Study the Structure of Extended Topic

$$P_{E-COHEN} = 0.4543$$

$$\hat{k} = 50.2\%$$

$$P_{E-FLEISS} = 0.4570$$

$$\hat{k} = 49.9\%$$

(3) Subject 2 vs. Subject 3

	<i>Physics Term</i>	<i>Linguistic Term</i>	<i>General Scientific Term</i>	<i>Total</i>
<i>Physics Term</i>	8	1	3	12
<i>Linguistic Term</i>	0	19	9	28
<i>General Scientific Term</i>	5	22	38	63
<i>Total</i>	13	40	50	103

$$P_A = 0.6311$$

$$P_{E-COHEN} = 0.4172$$

$$\hat{k} = 36.7\%$$

$$P_{E-FLEISS} = 0.4246$$

$$\hat{k} = 35.9\%$$

The average agreements of pairs of subjects are as follows.

$$Kappa_{average-cohen} = 42.43\%$$

$$Kappa_{average-Fleiss} = 41.6\%$$

C.7 TERM CLASSIFICATION RESULT

The following are 100 terms, of which 60 terms are classified as scientific research general, 27 terms are considered as specific to Linguistics research, 13 terms are considered as specific to Physics research.

Term	Category
development	General
design	General
dialogue	Linguistics
technique	General
system	General
set	General

Term	Category
corpus	Linguistics
name	Linguistics
architecture	General
motivation	General
translation	Linguistics
text	Linguistics

Appendix C – Study the Structure of Extended Topic

Term	Category
word	Linguistics
flow	Physics
information	Linguistics
language	Linguistics
data	General
evaluation	General
section	General
corpora	Linguistics
method	General
conclusion	General
example	General
problem	General
idea	General
lexicon	Linguistics
present	General
learning	Linguistics
class	General
classification	General
constraint	General
dependency	General
result	General
device	General
application	General
approach	General
experiment	General
model	General
state	General
overview	General
process	General
set-up	General
procedure	General
theory	General
structure	General
characteristic	General
measurement	General
analysis	General
data	General
wave	Physics
laser	Physics
substrate	Physics
acquisition	General
use	General
field	General
study	General
investigation	General
methodology	General

Term	Category
solution	General
silicon	Physics
algorithm	Linguistics
control	General
simulation	General
principle	General
velocity	Physics
mechanism	Physics
proof	General
parser	Linguistics
sensor	Physics
classifier	Linguistics
framework	General
speech	Linguistics
verb	Linguistics
learning	Linguistics
dependency	General
question	General
platform	Linguistics
training	Linguistics
formalism	Linguistics
construction	Linguistics
chart	General
number	General
view	General
recognition	General
extraction	Linguistics
answer	Linguistics
feature	General
location	Linguistics
case	Linguistics
segmentation	Linguistics
module	General
actuator	Physics
temperature	Physics
array	General
shape	Physics
vibration	Physics
retrieval	Linguistics
material	Physics
surface	General
size	General

TABLE C-9. TERM CLASSIFICATION RESULT

C.8 LISTS OF MOST FREQUENT GENERAL TERMS AT THE HEAD NOUN POSITION

Term	Frequency
Method	37
System	26
Approach	26
Model	20
Experiment	17
Result	12
Feature	8
Set	7
Design	7
Procedure	5
Overview	5

Term	Frequency
Architecture	5
Motivation	4
Application	4
Data	4
Proof	4
Technique	4
Study	3
Framework	3
Implementatio n	3
Class	3

TABLE C-10. MOST FREQUENT GENERAL TERMS IN 'CL+DESCRIBE'

Term	Frequency
Approach	53
Method	45
Result	41
Model	16
Conclusion	11
Evaluation	10
System	9
Section	7
Framework	7
Experiment	7

Term	Frequency
Technique	7
Overview	6
Example	6
Problem	6
Idea	5
Feature	5
Analysis	5
Design	4
Solution	4
Present	4

TABLE C-11. MOST FREQUENT GENERAL TERMS IN 'CL+PRESENT'

Term	Frequency
Result	21
Study	20
Method	18
Design	16
Model	14
Analysis	12
System	9
Approach	8
Investigation	7
Technique	7

Term	Frequency
Simulation	5
Development	5
Methodology	5
Solution	4
Control	3
Principle	3
Procedure	3
Process	3
Application	3
Overview	2

TABLE C-12. MOST FREQUENT GENERAL TERMS IN 'PHYSICS+DESCRIBE'

Term	Frequency
Method	20

Term	Frequency
Development	13

Appendix C – Study the Structure of Extended Topic

Term	Frequency
Design	12
Technique	7
System	6
Application	5
Operation	4
Device	3
Approach	3
Result	3
Experiment	3

Term	Frequency
Model	3
State	2
Process	2
Set-up	2
Procedure	2
Theory	2
Structure	2
Characteristic	2
Use	1

TABLE C-13. MOST FREQUENT GENERAL TERMS IN 'PHYSICS+PRESENT'

APPENDIX D

VERIFY THE RELATION BETWEEN DIFFERENT PARTS OF TOPIC AND DIFFERENT DISCOURSE CONSTITUENCIES

This appendix contains materials used in the experiment introduced in section 4.5. The purpose of this experiment is to verify the relation between different parts of topic and different discourse constituencies. Specifically, the generic part of a topic expression is related to the *new* element in a relevant discourse and the specific part could appear as the *given* element in the discourse. To verify the relation, we generate *signatures* for each part of a topic and then observe where they occur in the relevant discourses. The signatures of a topic are a list of terms that are shared among a set of documents about the topic. In section D.1 we introduce the documents that are used to generate signatures and to verify the aforementioned mapping relation; section D.2 gives examples of topic signatures.

D.1 EXPERIMENT MATERIAL

GROUP 1

CAUSES OF LUNG CANCER

Most lung cancers are caused by cigarette smoking. The more cigarettes you smoke per day and the earlier you started smoking, the greater the risk of lung cancer.

Second-hand smoke has also been shown to increase risk. Government surveys show that as many as 3,000 people each year develop lung cancer from second-hand smoke. High levels of pollution, radiation, and asbestos exposure may also increase risk.

Lung cancer begins by changes in cells that line the airways and can invade adjacent tissues before symptoms are noticed.

There are many types of lung cancer, but most can be categorized into two basic types, "small cell" and "non-small cell." Small cell lung cancer is generally faster growing than non-small cell, but more likely to respond to chemotherapy.

Small cell cancer is divided into "limited stage" (generally cancer confined to the chest) and "extensive stage" (cancer that has spread outside the chest).

Non-small cell cancer is divided into four stages, I-IV. Most patients with stage I and II non-small cell tumors and some patients with stage III tumors can undergo surgery with the goal of

*Appendix D – Verify the Relation between
Different Parts of Topic and Different Discourse Constituencies*

cure. Stage IV denotes cancer that has spread to other sites in the body (most often bone, brain, or liver) and is, in most cases, not curable.

Men and women die from lung cancer more than any other type of cancer. Lung cancer occurs most often in people between 55 and 65 years old.

PREVENTION OF LUNG CANCER

Lung cancer; Bronchogenic cancer; Cancer - lung (primary); Small cell lung cancer; Non-small cell lung cancer

If you smoke, stop smoking. Try to avoid second-hand smoke.

There is no conclusive evidence that screening for lung cancer with chest X-rays or CT scans is beneficial for patients at high risk of developing lung cancer. However, some recent studies have suggested that specialized CT scans called "spiral CT scans" may help improve cure rates by detecting lung cancer at an earlier stage. This is still under investigation.

SYMPTOMS OF LUNG CANCER

Lung cancer; Bronchogenic cancer; Cancer - lung (primary); Small cell lung cancer; Non-small cell lung cancer

- Cough
- Bloody sputum
- Shortness of breath
- Chest pain
- Loss of appetite
- Weight loss
- Additional symptoms that may be associated with this disease:
- Weakness
- Swallowing difficulty
- Nail abnormalities
- Joint pain
- Hoarseness or changing voice
- Fever
- Facial swelling
- Facial paralysis
- Eyelid drooping

*Appendix D – Verify the Relation between
Different Parts of Topic and Different Discourse Constituencies*

- Bone pain or tenderness

TREATMENT OF LUNG CANCER

Lung cancer; Bronchogenic cancer; Cancer - lung (primary); Small cell lung cancer; Non-small cell lung cancer

The treatment depends upon the type of cancer and the stage of the disease.

For small cell cancer, chemotherapy and radiation are usually used when the disease is confined to the chest -- so called "limited stage" disease. Chemotherapy alone is used in other situations (e.g., extensive stage disease).

For non-small cell cancer:

Surgical resection (cutting out the tumor) is usually done when the cancer has not spread beyond the lung and selected lymph nodes -- stage I, II and selected cases of stage III.

The combination of chemotherapy and radiation therapy is often used for cancer confined to the lung and lymph nodes that cannot be removed by surgery (stage III).

Some patients will undergo chemotherapy or a combination of chemotherapy and radiation prior to surgery.

Chemotherapy alone is used when the cancer is metastatic (stage IV); chemotherapy has been shown to prolong survival and improve quality of life.

GROUP 2

CLINICAL PHARMACOLOGY OF DIAZEPAM

In animals, diazepam appears to act on parts of the limbic system, the thalamus and hypothalamus, and induces calming effects. Diazepam, unlike chlorpromazine and reserpine, has no demonstrable peripheral autonomic blocking action, nor does it produce extrapyramidal side effects; however, animals treated with diazepam do have a transient ataxia at higher doses. Diazepam was found to have transient cardiovascular depressor effects in dogs. Long-term experiments in rats revealed no disturbances of endocrine function. Injections into animals have produced localized irritation of tissue surrounding injection sites and some thickening of veins after intravenous use.

DOSAGE OF DIAZEPAM

*Appendix D – Verify the Relation between
Different Parts of Topic and Different Discourse Constituencies*

Dosage should be individualized for maximum beneficial effect. The usual recommended dose in adults ranges from 2 mg to 20 mg IM or IV, depending on the indication and its severity. In some conditions, eg, tetanus, larger doses may be required. (See dosage for specific indications.) In acute conditions the injection may be repeated within 1 hour although an interval of 3 to 4 hours is usually satisfactory. Lower doses (usually 2 mg to 5 mg) and slow increase in dosage should be used for elderly or debilitated patients and when other sedative drugs are administered (see WARNINGS and ADVERSE REACTIONS).

For dosage in pediatric patients above the age of 30 days, see the specific indications below. When intravenous use is indicated, facilities for respiratory assistance should be readily available.

PRECAUTION OF DIAZEPAM

Although seizures may be brought under control promptly, a significant proportion of patients experience a return to seizure activity, presumably due to the short-lived effect of Valium after IV administration. The physician should be prepared to readminister the drug. However, Valium is not recommended for maintenance, and once seizures are brought under control, consideration should be given to the administration of agents useful in longer term control of seizures.

If Valium is to be combined with other psychotropic agents or anticonvulsant drugs, careful consideration should be given to the pharmacology of the agents to be employed-particularly with known compounds which may potentiate the action of Valium, such as phenothiazines, narcotics, barbiturates, MAO inhibitors and other antidepressants. In highly anxious patients with evidence of accompanying depression, particularly those who may have suicidal tendencies, protective measures may be necessary. The usual precautions in treating patients with impaired hepatic function should be observed. Metabolites of Valium are excreted by the kidney; to avoid their excess accumulation, caution should be exercised in the administration to patients with compromised kidney function.

Since an increase in cough reflex and laryngospasm may occur with peroral endoscopic procedures, the use of a topical anesthetic agent and the availability of necessary countermeasures are recommended.

Until additional information is available, diazepam injection is not recommended for obstetrical use.

*Appendix D – Verify the Relation between
Different Parts of Topic and Different Discourse Constituencies*

Valium Injection has produced hypotension or muscular weakness in some patients particularly when used with narcotics, barbiturates or alcohol.

Lower doses (usually 2 mg to 5 mg) should be used for elderly and debilitated patients.

The clearance of Valium and certain other benzodiazepines can be delayed in association with Tagamet (cimetidine) administration. The clinical significance of this is unclear.

SIDE EFFECT OF DIAZEPAM

Side effects most commonly reported were drowsiness, fatigue and ataxia; venous thrombosis and phlebitis at the site of injection. Other adverse reactions less frequently reported include: CNS: confusion, depression, dysarthria, headache, hypoactivity, slurred speech, syncope, tremor, vertigo. GI: constipation, nausea. GU: incontinence, changes in libido, urinary retention. Cardiovascular: bradycardia, cardiovascular collapse, hypotension. EENT: blurred vision, diplopia, nystagmus. Skin: urticaria, skin rash. Other: hiccups, changes in salivation, neutropenia, jaundice. Paradoxical reactions such as acute hyperexcited states, anxiety, hallucinations, increased muscle spasticity, insomnia, rage, sleep disturbances and stimulation have been reported; should these occur, use of the drug should be discontinued. Minor changes in EEG patterns, usually low-voltage fast activity, have been observed in patients during and after Valium therapy and are of no known significance.

In peroral endoscopic procedures, coughing, depressed respiration, dyspnea, hyperventilation, laryngospasm and pain in throat or chest have been reported.

Because of isolated reports of neutropenia and jaundice, periodic blood counts and liver function tests are advisable during long-term therapy.

GROUP 3

CLIMATE OF FIJI

At Suva the average summer high temperature is 85 F (29 C) and the average winter low is 68 F (20 C); temperatures typically are lower in elevated inland areas. All districts receive the greatest amount of rainfall in the season from November through March, during which time hurricanes are also experienced perhaps once every two years. While rainfall is reduced in the east of the larger islands from April to October, giving an annual average of 120 inches (3,050 millimetres) per year, it virtually ceases in the west, to give an annual rainfall of 70 inches, thus making for a sharp contrast in both climatic conditions and agriculture between east and west.

*Appendix D – Verify the Relation between
Different Parts of Topic and Different Discourse Constituencies*

CULTURE BACKGROUND OF FUJI

Fiji's mixed racial background contributes to a rich cultural heritage. Many features of traditional Fijian life survive; they are most evident in the elaborate investiture, marriage, and other ceremonies for high-ranking chiefs. These ceremonies provide a focus for the practicing of traditional crafts, such as the manufacture of masi, or tapa cloth, made from the bark of the paper mulberry; mat weaving; wood carving; and canoe making. Drinking of yanggona (kava, made from the root of *Piper methysticum*) is a part not only of important ceremonies but also of everyday life. Displays of traditional Fijian culture, music, and dancing make an important contribution to tourism; model villages and handicraft markets are popular.

Most Indian women continue to wear the sari together with traditional jewelry in gold and silver. Traditional marriage ceremonies are practiced, as are customs such as fire walking and ritual self-torture as part of important religious ceremonies. Cinemas showing imported Indian films are popular. Diwali, the Hindu Festival of the Lights, is celebrated every October and is a public holiday.

Fiji has two daily newspapers and a multilingual public radio broadcasting system. Videocassette players are common in the towns (many villages have no electricity).

Ethnic composition of Fiji

While the indigenous Fijian people are usually classified as Melanesian, they are larger in stature than Melanesians from Vanuatu, Solomon Islands, or New Guinea; their social and political organization is closer to that of Polynesia; and there has been a high level of intermarriage between Fijians from the Lau Islands of eastern Fiji and the neighbouring Polynesian islands of Tonga. Almost all indigenous Fijians are Christian, mostly Methodist and Roman Catholic.

Since World War II, indigenous Fijians have been outnumbered by Indians, most of whom are descendants of indentured labourers brought to work in the sugar industry. A few, particularly in commerce and the professions, are descended from free migrants. Most of the Indians are Hindus, though a significant number are Muslims.

There are also significant minorities of Europeans, part-Europeans, Chinese, and Pacific islanders from outside Fiji. In the last group are the Polynesian population of Rotuma and the Banabans, who were forced to leave Banaba after destruction in World War II made it uninhabitable. Many Banabans settled on Rabi Island in Fiji.

*Appendix D – Verify the Relation between
Different Parts of Topic and Different Discourse Constituencies*

English is the official language. The widely used Fijian language has many dialects; the one most commonly used is known as Bauan Fijian and comes from Bau, the district that enjoyed political supremacy at the advent of colonial rule. Most people speak at least two languages, including English and the language of their own racial community.

There is very little intermarriage between racial communities. The relative proportions of Fijians and Indians in the population have been changing in recent years because the Fijian birth rate is higher than the Indian and because of the accelerating outward migration of Indians, especially to Canada, Australia, and New Zealand. While Suva has a very mixed population, the sugar regions of Viti Levu and Vanua Levu have predominantly Indian populations. On the smaller islands, and in less-developed rural areas of the larger islands, Fijians live in traditional villages. The two largest urban centres are on Viti Levu: Suva, in the southeast, which has about one-fifth of the total population, and Lautoka in the northwest, which is the centre of the sugar industry and has a major port. Labasa on Vanua Levu is a centre for administration, services, and sugar production.

Economic resources of Fiji

The third major export earner, though well behind sugar and tourism, is gold, which is mined at Vatukoula in northern Viti Levu. Copper deposits are known to exist at Namosi, inland from Suva, but mining is not viable.

Since large-scale systematic planting of pine forests began in the 1960s, a timber industry has developed for domestic use and export. Development plans have emphasized the need to reduce dependence on imported food, especially rice, meat, fish, and poultry products. There is substantial hydroelectricity generation, but fuel remains a major import together with manufactured goods (many for resale to tourists), machinery, and food. Australia, New Zealand, Japan, and the United States are the major sources of imports.

D.2 TOPIC SIGNATURES

GROUP 1

Generic Topic	Signatures
cause	however, great, increased, long, begin, affect, cause, much, genetic, usually, risk, year, have, rare, people, factor, also, birth, incidence, call, chronic, become, from, process, between, man, develop, disorder, sex, 20, condition
prevention	exposure, disease, low, regular, early, prevention, have, physician,

*Appendix D – Verify the Relation between
Different Parts of Topic and Different Discourse Constituencies*

	age, people, prevent, reduce, preventable, symptom, stop, may, when, help, particularly, possible, avoid, include, blood, develop, recommend, risk, woman, smoke, factor, can
symptom	excessive, disease, have, breath, loss, symptom, associate, should, may, chest, change, inability, difficulty, fatigue, include, appetite, weight, blood, cancer, swallow, skin, pain, about, shortness, often
treatment	especially, low, out, reduce, place, treatment, include, use, surgical, often, medication, risk, have, also, case, form, mild, type, apply, fatigue, call, diarrhea, individual, may, treat, destroy, indicate, nearby, procedure, early, can

Specific Topic	Signatures
lung cancer	cell, cancer, have, may, chest, lung, bronchogenic, small, patient, primary

GROUP 2

Generic Topic	Signatures
pharmacology	Nervous, however, therapeutic, pressure, inhibit, know, human, subject, have, clinical, also, establish, peripheral, effect, suppression, only, brain, function, action, activity, normal, unlike, primarily, from, may, between, appear, cord, follow, inhibition, system, mechanism, evidence, indicate, heart, unknown, 4, 12
Indication dosage	Dose, achieve, adjust, than, reduction, administer, reduce, should, may, increase, range, require, recommended, according, usually, indicate, patient, over, hour, use, day, response, from, recommend, do, administration, dosage, can, maximum, give, 6, therapy, after, 12
precaution	Child, however, caution, potential, benefit, when, under, possibility, use, risk, impaired, have, drug, precaution, symptom, establish, significant, rat, should, safety, function, report, patient, outweigh, avoid, unknown
side-effect	headache, adverse, diarrhea, cardiovascular, drug, nausea, system, reaction, effect, libido, hydrochloride, side, following, much, use, mouth, insomnia, dizziness, constipation, rash, edema, gastrointestinal, nervous, urticaria, hypotension

Specific Topic	Signatures
diazepam	usually, diazepam, term, after, have, increase, injection, effect, low, long, use, drug, function

*Appendix D – Verify the Relation between
Different Parts of Topic and Different Discourse Constituencies*

GROUP 3

Generic Topic	Signatures
climate	Millimetre, low, great, region, range, during, inch, september, have, wind, receive, drop, climate, rain, temperature, average, annually, approximately, high, from, October, amount, about, annual, 80, snow, f, c, 27, 25
cultural	begin, also, centre, radio, use, popular, contemporary, cultural, life, modern, however, art, musical, have, among, form, between, from, play, include, first, major, place, much, tradition
people	one, large, region, about, linguistic, century, people, cultural, area, little, number, two, country, find, have, official, society, population, ethnic, 19 th , include, various, come, language, background, urban, major, religious, part
resource	Petroleum, however, coal, limestone, ore, zinc, metallic, gold, deposit, include, development, mining, use, hydroelectric, criteria, offshore, iron, mine, valuable, large, reserve, copper, only, product, thermal, become, small, lead, rich, mineral, develop, economic, country, resource, sea, important, nuclear, well

Specific Topic	Signatures
fiji	religious, Fiji, reduce, also, area, Hindu, know, district, mixed, use, high, especially, two, australia, tourism, imported, October, from, suva, gold, inland, background, together, zealand

APPENDIX E

COMPARISON BETWEEN GENERIC TOPIC AND SPECIFIC TOPIC

This appendix provides example passages to illustrate that a general concept could both work as a generic topic or as a specific topic. Given the two different roles, the content of a relevant discourse would be different. Below we provide examples for two concepts, biography and anatomy.

E.1 CONCEPT: BIOGRAPHY

When biography is a generic topic, a relevant discourse would contain biographical facts of a particular person (e.g., his education and career development); when it is a specific topic, a relevant document would talk about some general attributes of biography as a type of written work (e.g., key facts in a typical biography).

E.1.1 BIOGRAPHY AS A GENERIC TOPIC

Article 1: Charles Dickens biography
Source: http://www.dickens-literature.com/l_biography.html
Extended topic: 'the biography of Charles Dickens'
Excerpt: DICKENS, CHARLES JOHN HUFFAM (1812—1870), English novelist, was born on the 7th of February 1812 at a house in the Mile End Terrace, Commercial Road, Landport (Portsea)—a house which was opened as a Dickens Museum on 22nd July 1904. His father John Dickens (d. 1851), a clerk in the navy-pay office on a salary of £80 a year, and stationed for the time being at Portsmouth, had married in 1809 Elizabeth, daughter of Thomas Barrow, and she bore him a family of eight children, Charles being the second. In the winter of 1814 the family moved from Portsea in the snow, as he remembered, to London, and lodged for a time near the Middlesex hospital. The country of the novelist's childhood, however, was the kingdom of Kent, where the family was established in proximity to the dockyard at Chatham from 1816 to 1821. He looked upon himself in later years as a man of Kent, and his capital abode as that in Ordnance Terrace, or 18 St Mary's Place, Chatham,

Appendix E – Comparison between Generic Topic and Specific Topic

amid surroundings classified in Mr Pickwick's notes as " appearing "to be soldiers, sailors, Jews, chalk, shrimps, officers and dockyard men. He fell into a family the general tendency of which was to go down in the world, during one of its easier periods (John Dickens was now fifth clerk on £250 a year), and he always regarded himself as belonging by right to a comfortable, genteel, lower middleclass stratum of society. His mother taught him to read; to his father he appeared very early in the light of a young prodigy, and by him Charles was made to sit on a tall chair and warble popular ballads, or even to tell stories and anecdotes for the benefit of fellow-clerks in the office.

Article 2: Gandhi - Biography of Mahatma Gandhi

Source: <http://history1900s.about.com/od/people/a/gandhi.htm>

Extended topic: 'the biography of Mahatma Gandhi'

Excerpt:

Historical Importance:

Mohandas Gandhi is considered the father of the Indian independence movement. Gandhi spent twenty years in South Africa working to fight discrimination. It was there that he created his concept of satyagraha, a non-violent way of protesting against injustices. While in India, Gandhi's obvious virtue, simplistic lifestyle, and minimal dress endeared him to the people. He spent his remaining years working diligently to both remove British rule from India as well as to better the lives of India's poorest classes. Many civil rights leaders, including Martin Luther King Jr., used Gandhi's concept of non-violent protest as a model for their own struggles.

Dates:

October 2, 1869 - January 30, 1948

Also Known As:

Mohandas Karamchand Gandhi, Mahatma ("Great Soul"), Father of the Nation, Bapu ("Father"), Gandhiji

Overview of Gandhi:

Appendix E – Comparison between Generic Topic and Specific Topic

Mohandas Gandhi was the last child of his father (Karamchand Gandhi) and his father's fourth wife (Putlibai). During his youth, Mohandas Gandhi was shy, soft-spoken, and only a mediocre student at school. Although generally an obedient child, at one point Gandhi experimented with eating meat, smoking, and a small amount of stealing -- all of which he later regretted. At age 13, Gandhi married Kasturba (also spelled Kasturbai) in an arranged marriage. Kasturba bore Gandhi four sons and supported Gandhi's endeavors until her death in 1944.

Article 3: Xu Zhimo

Source: http://en.wikipedia.org/wiki/Xu_Zhimo

Extended topic: 'the biography of Xu, Zhimo'

Excerpt:

Xu Zhimo (Chinese: 徐志摩; pinyin: *Xú Zhì mó*; Wade–Giles: Hsü Chih-mo, January 15, 1897—November 19, 1931) was an early 20th century Chinese poet. He was given the name of Zhangxu (章垿) and the courtesy name of Yousen (標森). He later changed his courtesy name to Zhimo (志摩).

He is romanticized as pursuing love, freedom, and beauty all his life (from the words of Hu Shi). He promoted the form of modern Chinese poetry, and therefore made tremendous contributions to modern Chinese literature.

To commemorate Xu Zhimo, in July, 2008, a white marble stone has been installed at the back of King's College, University of Cambridge, on which is inscribed a verse from Xu's best-known poem, 'Saying Goodbye to Cambridge Again'.

Brief Biography

Xu was born in Haining, Zhejiang and graduated from the famous Hangzhou High School (浙江省杭州高级中学). In 1915, he married Zhang Youyi and next year he went to Peiyang University (Beiyang University, now Tianjin University) to study Law. In 1917, he transferred to Peking University due to the law department of Peiyang University merging

Appendix E – Comparison between Generic Topic and Specific Topic

into Peking University. In 1918, after studying at Peking University, he traveled to the United States to study history in Clark University. Shortly afterwards, he transferred to Columbia University in New York to study economics and politics in 1919. Finding the States "intolerable", he left in 1922 to study at King's College, Cambridge in England, where he fell in love with English romantic poetry like that of Keats and Shelley, and was also influenced by the French romantic and symbolist poets, some of whose works he translated into Chinese. In 1922 he went back to China and became a leader of the modern poetry movement. In 1923, he founded the Crescent Moon Society.

When the Bengali poet Rabindranath Tagore visited China, Xu Zhimo played the part of oral interpreter. Xu's literary ideology was mostly pro-western, and pro-vernacular. He was one of the first Chinese writers to successfully naturalize Western romantic forms into modern Chinese poetry. He worked as an editor and professor at several schools before dying in a plane crash on November 19, 1931 in Jinan, Shandong while flying from Nanjing to Beijing. He left behind four collections of verse and several volumes of translations from various languages.

E.1.2 BIOGRAPHY AS PART OF A SPECIFIC TOPIC

Article 1: How to write an interesting biography?
Source: http://homeworktips.about.com/od/biography/a/bio.htm
Extended topic: 'the skills in writing an interesting biography'
Excerpt: A biography is a written account of the series of events that make up a person's life. The first information you should gather in your research will include biographical details and facts. Basic details include: <ul style="list-style-type: none">• Date and place of birth and death• Family information• Lifetime accomplishments• Major events of life

Appendix E – Comparison between Generic Topic and Specific Topic

- Effects/impact on society, historical significance

While this information is necessary to your project, these dry facts, on their own, don't really make a very good biography. Once you've found these basics, you'll want to dig a little deeper.

You choose a certain person because you think he or she is interesting, so you certainly don't want to burden your paper with an inventory of boring facts. Your goal is to impress your reader!

You'll want to start off with great first sentence.

... ..

Now that you've created an impressive beginning, you'll want to continue the flow. Find more intriguing details about the man and his work, and weave them into the composition.

... ..

Questions to consider in your biography:

- Was there something in your subject's childhood that shaped his/her personality?
- Was there a personality trait that drove him/her to succeed or impeded his progress?
- What adjectives would you use to describe him/her?
- What were some turning points in this life?
- What was his/her impact on history?

Article 2: What makes a good biography or autobiography?

Source: <http://legacymultimedia.com/blog/2008/05/02/what-makes-a-good-biography-or-autobiography/>

Extended topic: 'the content of a good biography or autobiogprahy'

Excerpt:

It would be fair to say that the more talented, famous, accomplished or successful a person is, the more interesting their life story will be. However, everyone has a unique set of experiences that can make for a highly entertaining and worthwhile biography or autobiography.

Appendix E – Comparison between Generic Topic and Specific Topic

... ..

The particulars of a person’s childhood, family and friends, cultural background, religious beliefs, trials, tribulations and passions all define them. So too do their personal tastes, sense of humor, jobs, hobbies, travels, and their involvement in whatever community or communities they belonged to along the way. In short, no two people have the exact same story and when told with flair, any life can serve as a compelling biographical subject.

... ..

From there you can build a series of scenes around the chronological events of a person’s life that tell their story as it unfolded. To embellish the text, you can employ whimsy, wit, humor, romance, drama, irony and even tragedy as literary or cinematic devices. After all, everyone experiences all those things at one time or another and they can make the difference between a dry, boring read and a real page-turner.

Article 3: How to write a biography?

Source: <http://www.infoplease.com/homework/wsbiography.html>

Extended topic: ‘the method of writing a biography’

Excerpt:

A biography is simply the story of a life.

... ..

Biographies analyze and interpret the events in a person's life. They try to find connections, explain the meaning of unexpected actions or mysteries, and make arguments about the significance of the person's accomplishments or life activities. Biographies are usually about famous, or infamous people, but a biography of an ordinary person can tell us a lot about a particular time and place. They are often about historical figures, but they can also be about people still living.

Many biographies are written in chronological order. Some group time periods around a major theme (such as "early adversity" or "ambition and achievement"). Still others focus on specific topics or accomplishments.

... ..

To write a biography you should:

1. Select a person you are interested in
2. Find out the basic facts of the person's life. Start with the encyclopedia and almanac.
3. Think about what else you would like to know about the person, and what parts of the life you want to write most about. Some questions you might want to think about include:
 - What makes this person special or interesting?
 - What kind of effect did he or she have on the world? other people?
 - What are the adjectives you would most use to describe the person?
 - What examples from their life illustrate those qualities?
 - What events shaped or changed this person's life?
 - Did he or she overcome obstacles? Take risks? Get lucky?
 - Would the world be better or worse if this person hadn't lived? How and why?
4. Do additional research at your library or on the Internet to find information that helps you answer these questions and tell an interesting story.
5. Write your biography. See the Tips on Writing Essays and How to Write a Five Paragraph Essay for suggestions.

E.2 CONCEPT: ANATOMY

E.2.1 ANATOMY AS A GENERIC TOPIC

When anatomy is a generic topic, a relevant discourse would contain detailed explanation of the structure or composition of an organism (e.g., the anatomy of the jaw); when it is a specific topic, a relevant document would talk about some general attributes of anatomy as a scientific discipline (e.g., scope of the study of anatomy).

Article 1: Liver
Source: http://en.wikipedia.org/wiki/Liver#Anatomy
Extended topic: 'the anatomy of the liver'
Excerpt:

Appendix E – Comparison between Generic Topic and Specific Topic

The liver is a reddish brown organ with four lobes of unequal size and shape. A human liver normally weighs 1.44–1.66 kg (3.2–3.7 lb), and is a soft, pinkish-brown, triangular organ. It is both the largest internal organ (the skin being the largest organ overall) and the largest gland in the human. It is located in the right upper quadrant of the abdominal cavity, resting just below the diaphragm. The liver lies to the right of the stomach and overlies the gallbladder. It is connected to two large blood vessels, one called the hepatic artery and one called the portal vein. The hepatic artery carries blood from the aorta, whereas the portal vein carries blood containing digested nutrients from the entire gastrointestinal tract and also from the spleen and pancreas. These blood vessels subdivide into capillaries, which then lead to a lobule. Each lobule is made up of millions of hepatic cells which are the basic metabolic cells.

Article 2: Anatomy of the Brain

Source: <http://biology.about.com/od/humananatomybiology/a/anatomybrain.htm>

Extended topic: 'the anatomy of the brain'

Excerpt:

The anatomy of the brain is complex due its intricate structure and function. This amazing organ acts as a control center by receiving, interpreting, and directing sensory information throughout the body. There are three major divisions of the brain. They are the forebrain, the midbrain, and the hindbrain.

Anatomy of the Brain: Brain Divisions

The forebrain is responsible for a variety of functions including receiving and processing sensory information, thinking, perceiving, producing and understanding language, and controlling motor function. There are two major divisions of forebrain: the diencephalon and the telencephalon. The diencephalon contains structures such as the thalamus and hypothalamus which are responsible for such functions as motor control, relaying sensory information, and controlling autonomic functions. The telencephalon contains the largest part of the brain, the cerebrum. Most of the actual information processing in the brain takes place in the cerebral cortex.

Appendix E – Comparison between Generic Topic and Specific Topic

The midbrain and the hindbrain together make up the brainstem. The midbrain is the portion of the brainstem that connects the hindbrain and the forebrain. This region of the brain is involved in auditory and visual responses as well as motor function.

The hindbrain extends from the spinal cord and is composed of the metencephalon and myelencephalon. The metencephalon contains structures such as the pons and cerebellum. These regions assist in maintaining balance and equilibrium, movement coordination, and the conduction of sensory information. The myelencephalon is composed of the medulla oblongata which is responsible for controlling such autonomic functions as breathing, heart rate, and digestion.

Article 3: Knee Anatomy

Source: <http://www.sportsinjuryclinic.net/cybertherapist/kneeanatomy.php>

Extended topic: 'the anatomy of the knee'

Excerpt:

Introduction to knee joint anatomy

The knee joint is the largest joint in the body, consisting of 4 bones and an extensive network of ligaments and muscles. Injuries to the knee joint are amongst the most common in sporting activities and understanding the anatomy of the joint is fundamental in understanding any subsequent pathology.

Bones of the knee joint

The knee is made up of four main bones- the femur (thigh bone), the tibia (shin bone), fibula (outer shin bone) and patella (kneecap). The main movements of the knee joint occur between the femur, patella and tibia. Each are covered in articular cartilage which is an extremely hard, smooth substance designed to decrease the frictional forces as movement occurs between the bones. The patella lies in an indentation at the lower end of the femur known as the intercondylar groove. At the outer surface of the tibia lies the fibula, a long thin bone that travels right down to the ankle joint.

The knee joint capsule

Appendix E – Comparison between Generic Topic and Specific Topic

The joint capsule is a thick ligamentous structure that surrounds the entire knee. Inside this capsule is a specialized membrane known as the synovial membrane which provides nourishment to all the surrounding structures. Other structures include the infrapatellar fat pad and bursa which function as cushions to exterior forces on the knee. The capsule itself is strengthened by the surrounding ligaments.

... ..

E.2.2 ANATOMY AS PART OF A SPECIFIC TOPIC

Article 1: Outline of human anatomy
Source: http://en.wikipedia.org/wiki/Outline_of_human_anatomy
Extended topic: ‘an outline of human anatomy’
Excerpt: Human anatomy, a branch of anatomy, is the scientific study of the morphology of the adult human. It is subdivided into gross anatomy and microscopic anatomy. Gross anatomy (also called topographical anatomy, regional anatomy, or anthropotomy) is the study of anatomical structures that can be seen by unaided vision. Microscopic anatomy is the study of minute anatomical structures assisted with microscopes, and includes histology (the study of the organization of tissues), and cytology (the study of cells).

Article 2: Anatomy - Definition
Source: http://www.wordiq.com/definition/Anatomy
Extended topic: ‘a definition of anatomy’
Excerpt: Anatomy (from the Greek anatome, from ana-temnein, to cut up), is the branch of biology that deals with the structure and organization of living things; thus there is animal anatomy (zootomy) and plant anatomy (phytonomy). The major branches of anatomy include comparative anatomy and human anatomy.

Appendix E – Comparison between Generic Topic and Specific Topic

Animal anatomy may include the study of the structure of different animals, when it is called comparative anatomy or animal morphology, or it may be limited to one animal only, in which case it is spoken of as special anatomy.

From a utilitarian point of view the study of humans is the most important division of special anatomy, and this human anatomy may be approached from different points of view. From that of Medicine it consists of a knowledge of the exact form, position, size and relationship of the various structures of the healthy human body, and to this study the term descriptive or topographical human anatomy is given, though it is often, less happily, spoken of as anthropotomy.

... ..

From the morphological point of view, however, human anatomy is a scientific and fascinating study, having for its object the discovery of the causes which have brought about the existing structure of humans, and needing a knowledge of the allied sciences of embryology or developmental biology, phylogeny, and histology.

Article 3: History of anatomy

Source: http://en.wikipedia.org/wiki/History_of_anatomy

Extended topic: 'the history of anatomy'

Excerpt:

The development of anatomy as a science extends from the earliest examinations of sacrificial victims to the sophisticated analyses of the body performed by modern scientists. It has been characterized, over time, by a continually developing understanding of the functions of organs and structures in the body. The field of Human Anatomy has a prestigious history, and is considered to be the most prominent of the biological sciences of the 19th and early 20th centuries. Methods have also improved dramatically, advancing from examination of animals through dissection of cadavers to technologically complex techniques developed in the 20th century.

Anatomy is one of the cornerstones of a doctor's medical education. Despite being a persistent portion of teaching from at least the renaissance, the format and the amount of

Appendix E – Comparison between Generic Topic and Specific Topic

information being taught has evolved and changed along with the demands of the profession. What is being taught today may differ in content significantly from the past but the methods used to teach this have not really changed that much. For example all the famous public dissections of the Middle Ages and early renaissance were in fact prosections. Prosection is the direction in which many current medical schools are heading in order to aid the teaching of anatomy and some argue that dissection is better. However looking at results of post graduate exams, medical schools (specifically Birmingham) that use prosection as opposed to dissection do very well in these examinations. This would suggest that prosection can fit very well into the structure of modern medical training.

APPENDIX F

ANSWERING CAUSAL QUESTIONS

This appendix includes materials related to the experiment on answering causal questions introduced in chapter 7. Section F.1 includes the causal indicators and examples of how they are used to indicate causal relation; section F.2 includes the patterns we use to match against a sentence to detect causal relation; section F.3 provides the sample documents we use in the document retrieval experiment.

F.1 CAUSAL INDICATORS AND USAGE EXAMPLES

Causal Indicator	Usage Example
'caused by'	Eight patients suffered from a frozen shoulder postoperatively, caused by a foreign body reaction and synovitis, and required revision surgery.
'induced by'	Sudden ankle inversion was induced by pulling the inversion platform support, allowing the platform support base to rotate 37 degrees.
'provoked by'	Chronic pain can be provoked by light contact stimuli.
'evoked by'	RAGE-mediated neutrophil dysfunction is evoked by advanced glycation end products.
'triggered by'	Malignant hyperthermia is triggered by volatile anaesthetics and/or depolarizing muscle relaxants by an abnormal increase of intracellular calcium concentration in skeletal muscle cells.
'stimulated by'	Atherosclerosis is an inflammatory disease stimulated by various infectious agents.
'cause'	Latex is composed of compounds that may cause an allergic reaction.
'induce'	Stimulation of the cells with EGF concomitantly induced an increase in intracellular Ca(2+), activation of CaMKII, and dissociation of PP2A-IQGAP1-CaMKII from beta1 integrin-Rac.
'provoke'	Fungal spore concentrations may provoke asthma attacks resulting in visits to A&E departments and emergency admission to hospital.

Appendix F – Answer Causal Questions

‘evoke’	To delineate the underlying mechanisms of hypertension-induced nephropathy, we generated transgenic mice that overexpress rat ANG (rANG) in the kidney to establish whether intrarenal RAS activation alone can evoke hypertension and kidney damage and whether RAS blockade can reverse these effects.
‘trigger’	Burns cause thermal injury to local tissue and trigger systemic acute inflammatory processes, which may lead to multiple distant organ dysfunction.
‘cause of’	Depression is another cause of psychogenic impotence.
‘causal agent of’	The HIV-1 envelope glycoprotein gp120 has been suggested to be a causal agent of neuronal loss.
‘causative agent of’	Ehrlichia ruminantium is the causative agent of heartwater, a major tick-borne disease of livestock in Africa that has been introduced in the Caribbean and is threatening to emerge and spread on the American mainland.
‘due to’	However, the subcategory of puncture wounds due to impalement by foreign bodies is quite rare.
‘stem from’	Autism may stem from problems in brain links.
‘arise from’	We have thus examined whether cervical DNA contains alkylation damage arising from exposure to methylating agents (N7-methyldeoxyguanosine, N7-MedG).
‘originate from’	Santavy will use a technique known as transitional electron microscopy to determine if the disease originates from a fungus, bacterium or virus.
‘result from’	Stress fractures are thought to result from a variety of causes, including muscular fatigue, sudden changes in training intensity or duration, and microtrauma to bone at the muscular origin and insertion sites ("wear-and-tear" theory).
‘result in’	This may result in testicular shrinkage (atrophy) and infertility.
‘leads to’/‘leading to’	Severe sepsis is frequently associated with adrenal insufficiency, which may lead to hemodynamic instability and a poor prognosis.
‘led to’	CCl(4)-induced cirrhosis in rats led to prolonged oxidative stress in the intestine.

Appendix F – Answer Causal Questions

'increase the risk of'	Research in monkeys suggests that a diet high in the natural plant estrogens found in soy does not increase the risk of breast or uterine cancer.
'increase the risk for'	Does Job Strain Increase the Risk for Coronary Heart Disease or Death in Men and Women.
'in the pathogenesis of'	Essential role of TNF family molecule LIGHT as a cytokine in the pathogenesis of hepatitis.
'as the result of'	Stress fractures occur as the result of increased remodeling and a subsequent weakening of the outer surface of the bone.
'as a ... result of'	Parathyroid excess may happen as a result of autonomous overactivity of the parathyroid glands (primary hyperparathyroidism).
'as the consequence of'	B/W mice develop renal disease as the consequence of a profound immunological perturbation.
'is a consequence of'	Rheumatic Valve Disease is a consequence of rheumatic fever.
'is the consequence of'	Pulmonary (lung) fibrosis, is a scarring of the lungs, and is the consequence of untreated pulmonary inflammation (alveolitis).
'are the consequence of'	Our results may be interpreted in line with the traditional view that allergy and atopic disease are the consequence of contact with antigen.
'are the effect of'	The classic clinical symptoms of allergic conjunctivitis (type I allergy) - itching and lacrimation - are the effect of histamine.
'is the result of'	Type I, insulin-dependent diabetes (IDDM), which becomes manifest before the age of 40, is the result of an absolute deficiency of insulin.
'are the result of'	The MMR-autism theory is based on the idea that intestinal problems, like Crohn's disease, are the result of viral infection.
'is a ... result of'	It is regarded that dementia in Alzheimer's disease is a result of the loss of neurons.

TABLE F-1. CAUSAL INDICATORS AND USAGE EXAMPLES

F.2 PATTERNS OF CAUSAL INDICATORS

Group	Causal Indicator	Pattern	Max
1	'caused by'	<causee>(+<wordstem>){0..Max}+caus+by	10/52
	'induced by'	<causee>(+<wordstem>){0..Max}+induc+by	
	'provoked by'	<causee>(+<wordstem>){0..Max}+provok+by	
	'evoked by'	<causee>(+<wordstem>){0..Max}+evok+by	
	'triggered by'	<causee>(+<wordstem>){0..Max}+trigger+by	
	'stimulated by'	<causee>(+<wordstem>){0..Max}+stimul+by	
2	'cause'	caus(+<wordstem>){0..Max}+<causee>	4/23
	'induce'	induc(+<wordstem>){0..Max}+<causee>	
	'provoke'	provok (+<wordstem>){0..Max}+<causee>	
	'evoke'	evok(+<wordstem>){0..Max}+<causee>	
	'trigger'	trigger(+<wordstem>){0..Max}+<causee>	
3	'cause of'	caus+of(+<wordstem>){0..Max}+<causee>	3/10
	'causal agent of'	causal+agent+of(+<wordstem>){0..Max}+<causee>	
	'causative agent of'	caus+agent+of(+<wordstem>){0..Max}+<causee>	
4	'due to'	<causee>(+<wordstem>){0..Max}+due+to	5/27
5	'stem from'	<causee>(+<wordstem>){0..Max}+stem+from	7/17
	'arise from'	<causee>(+<wordstem>){0..Max}+aris+from	
	'originate from'	<causee>(+<wordstem>){0..Max}+origin+from	
	'result from'	<causee>(+<wordstem>){0..Max}+result+from	
6	'result in'	result+in(+<wordstem>){0..Max}+<causee>	7/18
	'leads to'/'leading to'	lead+to (+<wordstem>){0..Max}+<causee>	
	'led to'	Led+to (+<wordstem>){0..Max}+<causee>	

Appendix F – Answer Causal Questions

	'increase the risk of'	increas+the+risk+of (+<wordstem>){0..Max}+<causee>	
7	'increase the risk for'	increas+the+risk+for (+<wordstem>){0..Max}+<causee>	2/5
	'in the pathogenesis of'	in+the+pathogenesi+of(+<wordstem>){0..Max}+<causee>	
8	'as the result of'	<causee>(+<wordstem>){0..Max}+as+the+result+of	1/2
9	'as a ... result of'	<causee>(+<wordstem>){0..Max}+as+a(+<wordstem>){0,2}+result+in	5/20
	'as the consequence of'	<causee>(+<wordstem>){0..Max}+as+the+consequ+of	
	'is a consequence of'	<causee>(+<wordstem>){0..Max}+as+the+result+of	
10	'is the consequence of'	<causee>(+<wordstem>){0..Max}+is+the+consequ+of	3/12
	'are the consequence of'	<causee>(+<wordstem>){0..Max}+are+the+consequ+of	
	'are the effect of'	<causee>(+<wordstem>){0..Max}+are+the+effect+of	
	'is the result of'	<causee>(+<wordstem>){0..Max}+is+the+result+of	
	'are the result of'	<causee>(+<wordstem>){0..Max}+are+the+result+of	
	'is a ... result of'	<causee>(+<wordstem>){0..Max}+is+a(+<wordstem>){0,2}+result+of	

TABLE F-2. PATTERNS OF CAUSAL INDICATORS

F.3 SAMPLE EXPERIMENT TEXT

A sample document from the Muchmore project
<p>Small-fiber neuropathy: answering the burning questions.</p> <p>Fink E, Oaklander AL.</p> <p>Nerve Injury Unit, Departments of Anesthesiology, Neurology, and Neuropathology, Massachusetts General Hospital, Harvard Medical School, Boston, MA 02114, USA.</p>

Appendix F – Answer Causal Questions

efink@partners.org.

Small-fiber neuropathy is a peripheral nerve disease that most commonly presents in middle-aged and older people, who develop burning pain in their feet. Although it can be caused by disorders of metabolism such as diabetes, chronic infections (such as with human immunodeficiency virus), genetic abnormalities, toxicity from various drugs, and autoimmune diseases, the cause often remains a mystery because standard electrophysiologic tests for nerve injury do not detect small-fiber function. Inadequate ability to test for and diagnose small-fiber neuropathies has impeded patient care and research, but new tools offer promise. Infrequently, the underlying cause of small-fiber dysfunction is identified and disease-modifying therapy can be instituted. More commonly, the treatments for small-fiber neuropathy involve symptomatic treatment of neuropathic pain.

Appendix F – Answer Causal Questions

A sample Web document from Pub Med

Source: <http://www.ncbi.nlm.nih.gov/pubmed/16522468>

NCBI Resources ▾ How To ▾

PubMed.gov PubMed ▾

US National Library of Medicine
National Institutes of Health [Limits](#) [Advanced](#)

[Display Settings:](#) (v) Abstract [Send to:](#) (v)

[J Allergy Clin Immunol. 2006 Mar;117\(3\):663-9. Epub 2006 Jan 27.](#)

Is occupational asthma to diisocyanates a non-IgE-mediated disease?

[Jones MG](#), [Floyd A](#), [Nouri-Aria KT](#), [Jacobson MR](#), [Durham SR](#), [Taylor AN](#), [Cullinan P](#).
Department of Occupational and Environmental Medicine, National Heart and Lung Institute, Faculty of Medicine, Imperial College, London, UK. meinir.jones@imperial.ac.uk

Abstract

BACKGROUND: Exposure to diisocyanates in the workplace is an important cause of occupational asthma. The majority of patients with diisocyanate-induced asthma have no detectable diisocyanate-specific IgE antibodies in serum. There has been much debate as to whether this is due to diisocyanate-induced asthma being mediated by non-IgE mechanisms or whether it is the result of using inappropriate conjugates.

OBJECTIVE: We sought to determine whether RNA message for Cepsilon, IL-4, and other associated inflammatory markers could be detected locally within the bronchial mucosa after diisocyanate challenge.

METHODS: Fiberoptic bronchoscopic bronchial biopsy specimens were obtained at 24 hours after both a control and an active challenge in 5 patients with positive and 7 patients with negative inhalation test responses to diisocyanates. Using both immunohistochemistry and in situ hybridization, we determined mRNA for Cepsilon, IL-4, IL-5, and other associated inflammatory markers.

RESULTS: There was a striking absence of Cepsilon and IL-4 mRNA-positive cells in bronchial biopsy specimens from patients challenged with diisocyanate (Cepsilon median of 0 and interquartile range of 0-1.85; IL-4 median of 0 and interquartile range of 0-0.85). In contrast, there were increased numbers of IL-5-, CD25-, and CD4-positive cells and a trend toward an increase in eosinophils after active challenge with diisocyanate.

CONCLUSION: We found a striking absence of both bronchial Cepsilon and IL-4 RNA message after inhalation challenge with diisocyanates, irrespective of whether the challenge test response was positive or negative. We propose that diisocyanate-induced asthma is a non-IgE-mediated disease, at least in patients in whom specific IgE antibodies to diisocyanates are undetectable.

Comment in
[J Allergy Clin Immunol. 2007 Mar;119\(3\):757-8; author reply 758.](#)

PMID: 16522468 [PubMed - indexed for MEDLINE]

[+](#) **Publication Types, MeSH Terms, Substances**

[+](#) **LinkOut - more resources**

APPENDIX G

ANSWERING PROCEDURAL QUESTIONS

This appendix includes materials being used in the experiments on answering procedural questions; these experiments are introduced in section 8.2. Section G.1 provides the feature sets used to classify procedural text; section G.2 provides sample documents used in the classification and retrieval experiments.

G.1 PROCEDURAL FEATURE SET

Cue phrase	Morphological tag	Syntactic tag
After	%<E	A
Again	%>N	ABBR
Already	%AUX	ACC
As Long As	%CC	ADV
As soon as	%CS	AUXMOD
At once	%E>	CARD
Because	%EH	CMP
Before	%N<	DEM
Begin	%NH	GEN
By	%VA	Heur
Date	%VP	IMP
Else	@+FAUXV	INF
Finally	@+FMAINV	ING
First	@-FAUXV	N
Following	@-FMAINV	NEG-PART
for	@<NOM	ORD
If	@<NOM-OF	PAST
In order to	@<P	PERS
In the end	@<P-FMAINV	PL
Later	@A>	PL1
Move on	@APP	PL3
Must	@DN>	PRES
Next	@DUMMY	PRON

Appendix G – Answer Procedural Questions

Cue phrase	Morphological tag	Syntactic tag
Next step	@F-SUBJ	RECIPR
Next time	@I-OBJ	SG
Now	@INFMARK>	SG1
Now that	@NH	SG3
Once	@O-ADVL	SUBJUNCTIVE
Or	@OBJ	SUP
Should	@PCOMPL-O	V
Since	@PCOMPL-S	WH
So that	@QN>	<Interr>
Start	@SUBJ	<Refl>
Then	@VOC	<Rel>
This time		
Thus		
Time		
To		
To begin with		
Unless		
Until		
When		
Will		
You		

TABLE G-1. FULL PROCEDURAL FEATURE SET

Cue phrase	Morphological tag	Syntactic tag
After	%CS	ABBR
Again	@I-OBJ	ACC
Already	@INFMARK>	AUXMOD
As soon as	@NH	GEN
Before	@O-ADVL	Heur
Begin	@PCOMPL-O	IMP
Date		INF
Finally		NEG-PART
If		PERS

Appendix G – Answer Procedural Questions

In order to	PL3
Next	SUBJUNCTIVE
Now	WH
Once	<Rel>
Should	
So that	
Start	
Then	
Time	
To	
Until	
When	
Will	
You	

TABLE G-2. SELECTED DISTINCTIVE INDIVIDUAL FEATURE SET

G.2 SAMPLE EXPERIMENT TEXT

Sample text from the SPIRIT collection
<pre><?xml version="1.0" encoding="iso-8859-1" ?> <DOCUMENT>PhotoVoyage The Day in Photos Top Story News Video/Audio The Week in Review On Assignment AQ Photo Voyages On the Lightbox Best of the Post Photos From: OnPolitics Nation World Metro Business/Tech Sports Live Online Style Entertainment Education Travel Health FAQs Tools Resources Contact Us Related Links Travel stories Back to Photovoyage front Grand Tetons: Impressions of a Wild Frontier Grand Teton National Park, Wyoming: 310,000 acres, home to an awe inspiring array of wildlife and host to nearly 3.5 million tourists annually. Located in the northwest corner of Wyoming, below Yellowstone, the Teton Range stretches along a 40-mile-long fault beneath the Rockies. New Caledonia: French Destination Unknown New Caledonia is known as the "Paris of the Pacific", hosting chic French boutiques and fine French cuisine. A struggle for independence from France has been an ongoing theme for years, climaxing in 1980 and a resolve to take place in 2014. Inner Japan: Essence of Life Take a visual journey through Japan, see exotic gardens with thousand year old pagodas, carp fish bred to imitate "living</pre>

Appendix G – Answer Procedural Questions

flowers", and be a spectator to a country full of festivals celebrating people, land and life.

Ireland's Aran Islands: The isolated Aran Islands stand off of the coast of Ireland in the Atlantic, a symbol of solitude and tradition. Tourists, trade and a dwindling population are now challenging the face of this unique place that didn't have electricity until the 1970's.

The Basque: First China's Silk Road Since the days of Marco Polo - China's Silk Road, a series of ancient trade routes criss-crossing Asia, has conjured up images of mystery, intrigue, and adventure. Today, camel caravans no longer travel the road yet vibrant marketplaces from Kashkar to Beijing continue to conduct the timeless trade of business - in both traditional and modern forms.</DOCUMENT>

Sample pagewise document

Source: <http://www.essortment.com/paint-mural-60710.html>



How to paint a mural

Learn how to paint a mural! Some simple and concise directions that will lead any beginning painter through some easy steps toward painting a mural!

Sponsored Links

- Mosaic & Fresco Class**
Learn to make mosaics & fresco wall murals on this mini-course in Italy
www.osmaonline.org
- Outstanding Murals**
Mural and Trompe l'oeil painting Beautiful patinas too!
www.grahamsmenage.com
- Fine Portrait Miniatures**
Antique, intimate and exquisite Hold history in your hand
www.ArchiebaldMiniatures.com
- Incredible Results**
But Not On Your First Try Free 10-Part Video Explains Why
www.virtualartacademy.com

Ads by Google

Here are some easy steps to follow in order to have a wonderful mural in your home! A mural will personalize your home or room and will leave a lasting impression on guests and family. You can involve your friends or loved-ones in the production of the mural, furthering an intimate piece of art that reflects your individuality!

Suggested places for a mural: your child's room, the [bathroom](#), a hallway, the [kitchen](#), the game room.

Suggested images: landscape, psychedelic flowers, fairies, castle, butterflies, fruit, cartoon characters, ANYTHING THAT YOU CAN DRAW! Start with something simple.

Ads by Google

- [Easter Craft](#)
- [Kids Craft](#)
- [Crafts Ideas](#)
- [Craft Project](#)

Supplies that you will need:

1. a fairly smooth wall on which to paint, with a flat finish (glossy won't work)
 2. pencils
 3. [stencils](#) (if you are timid about drawing freehand)
 4. brightly-colored acrylic [craft](#) or flat house paints
 5. BIG drop cloth
 6. a variety of [paint brush](#) sizes (very small for detail to 4 inch for painting large surfaces)
 7. [containers](#) to hold rinse-water for brushes
- Note: It is much easier to [paint](#) on a light-colored surface. If the wall you choose is a dark color (navy, dark green, etc.) paint it white or some other light color before you begin the project.

Steps to creating your mural:

1. Clear the area surrounding the mural wall of all furniture, toys, etc.
2. Lay down the [drop cloth](#), making sure to protect all carpeted area.
3. Use your pencil to draw the image directly onto the wall, stepping back several feet periodically to make sure the proportions of your images are correct. Don't forget to DRAW BIG SHAPES!!!
4. Fill in the shapes with the colors of your choice, starting with the background images and layering the foreground images on top.
5. Outline the foreground images with darker colors in order to emphasize edges.
6. Step back as far as you can and look for obvious errors.
7. Correct the errors using light colored paint and then re-draw and re-paint.
8. Step back again and view the final image.
9. When completely dry to the touch, clean up the area and return furniture, etc. the proper place.

Mosaic & Fresco Class
Learn to make mosaics & fresco wall murals on this mini-course in Italy
www.osmaonline.org

Fine Portrait Miniatures
Antique, intimate and exquisite Hold history in your hand.
www.ArchiebaldMiniatures.com

Top Rated Anti-Graffiti
Spray Paint & Marker Repellent... Water Based, Environment Friendly
www.vearforcorp.com

Ads by Google

© 2002 Pagewise

Tell us what topic you want to read about.

Request a new article

Submit

You are here: [Essortment Home](#) >> [Arts & Entertainment](#) >> [Art:Painting](#) >> [How to paint a mural](#)

<<[The nature of women and art](#)

[Painting with pastels](#)>>

DISCLAIMER - PLEASE READ - By printing, downloading, or using you agree to our full terms. Review the full terms at the following URL: <http://www.pagewise.com/diskclaimer.htm>. Below is a summary of some of the terms. If you do not agree to the full terms, do not use the information. We are only publishers of this material, not authors. Information may have errors or be outdated. Some information is from historical sources or represents opinions of the author. It is for research purposes only. The information is "AS IS", "WITH ALL FAULTS". User assumes all risk of use, damage, or injury. You agree that we have no liability for any damages. We are not liable for any consequential, incidental, indirect, or special damages. You indemnify us for claims caused by you.

FAQs: This site is published by [Pagewise, Inc.](#) Would you like to [link](#) to this page? Reprint this article on your [website](#)? Reprint this article on [pages](#)? Want to [reference](#) this article in a paper, report, or presentation? Is there an [error](#) in this page? Do you have a follow-up [question](#) about this topic? Want to read our [Privacy Policy](#)? Read our legal/medical [disclaimers](#)?



Appendix G – Answer Procedural Questions

A sample Web document found by using Google search

Source: http://www.s4c.co.uk/dudley/rm/view_recipe/rid/386/language/eng/

Dudley

RECIPE SEARCH SEARCH CYMRAEG

HOMEPAGE

RECIPES

> Categories

> This week's recipes


TASTY BITS

> All the tasty bits...

> Welsh beef

> Special diets

> Seafood



NEFYN HERRING WITH HOT AND SOUR SAUCES AND SALSA VERDE

CATEGORY : FISH & SHELLFISH

ENOUGH FOR : 3-6

INGREDIENTS

6 herrings, ready filleted

The hot marinade:

1 teaspoon of olive oil
1 small onion, finely chopped
2 to 3 teaspoons of Cajun spices
1 tablespoon of red or white wine
1 tablespoon of tomato puree

The sweet marinade

1 lemon, juice and zest
2 tablespoons of sherry
1 tablespoon of brown sugar
1 tablespoon of chives

The salsa verde:

3 tablespoons of fresh mint, finely chopped
3 tablespoons of fresh coriander, finely chopped
2 tablespoons of chives, finely chopped
½ a red chili, seeded
2 teaspoons of sesame oil
2 tablespoons of vegetable oil
1 tablespoon of clear honey
1 lemon, juice only

METHOD

Cut the fillets in 2 inch pieces and place on skewers, 2 to 3 for each skewer.

The hot marinade:
Heat the oil in a saucepan.

Cook the onions until soft and then add all the other ingredients.

Cook for 5 minutes then leave to cool.

The sweet marinade
Put the zest and the sugar in a pestle and mortar and grind together.

Put in the rest of the ingredients and mix until a smooth paste.

Divide the skewers between 2 dishes and pour the hot marinade over one and the sweet marinade over the other and leave to marinate for an hour.

The salsa verde:
Grind the herbs finely in a pestle and mortar and combine all the ingredients to create the salsa.

Cook the herring under a grill or on a barbecue for 10 minutes and serve with the salsa.

PROGRAMMES

> Gower to the Severn

> Tenby to Ferryside

> Fishguard to Milford Haven

> Aberaeron to Pwllgwaelod

> Portmeirion to Aberystwyth

> Llyn Peninsula

> Anglesey to Felinheli

> Deeside to Conway

PROGRAMME ARCHIVE

> 2006

> 2005

> 2004

EXTRAS

> Kitchen Essentials


> Food Suppliers

> Kitchenware Suppliers


> Ingredients

> Cooking Terms

> Dudley's Storecupboard



print recipe



download and print recipe
in schools format

DUDLEY
A Teledu Opus production for S4C

APPENDIX H

ANSWERING BIOGRAPHICAL QUESTIONS

This appendix includes materials being used in the experiments on answering biographical; these experiments are introduced in section 8.3. Section H.1 provides sample documents used in the classification and retrieval experiments; section H.2 provides examples of features that are applied to identify texts containing biographical facts.

H.1 SAMPLE BIOGRAPHIES

Sample text from Reuter
<pre><REUTERS TOPICS="NO" LEWISSPLIT="TRAIN" CGISPLIT="TRAINING-SET" OLDID="5547" NEWID="4"> <TEXT><TITLE>TALKING POINT/BANKAMERICA &It;BAC> EQUITY OFFER</TITLE> <AUTHOR> by Janie Gabbett, Reuters</AUTHOR> <DATELINE> LOS ANGELES, Feb 26 - </DATELINE> <BODY>BankAmerica Corp is not under pressure to act quickly on its proposed equity offering and would do well to delay it because of the stock's recent poor performance, banking analysts said. Some analysts said they have recommended BankAmerica delay its up to one-billion-dlr equity offering, which has yet to be approved by the Securities and Exchange Commission. BankAmerica stock fell this week, along with other banking issues, on the news that Brazil has suspended interest payments on a large portion of its foreign debt. The stock traded around 12, down 1/8, this afternoon, after falling to 11-1/2 earlier this week on the news. Banking analysts said that with the immediate threat of the First Interstate Bancorp &It;I> takeover bid gone, BankAmerica is under no pressure to sell the securities into a market that will be nervous on bank stocks in the near term. BankAmerica filed the offer on January 26. It was seen as one of the major factors leading the First Interstate withdrawing its takeover bid on February 9.</pre>

Appendix H – Answer Biographical Questions

A BankAmerica spokesman said SEC approval is taking longer than expected and market conditions must now be re-evaluated.

"The circumstances at the time will determine what we do," said Arthur Miller, BankAmerica's Vice President for Financial Communications, when asked if BankAmerica would proceed with the offer immediately after it receives SEC approval.

"I'd put it off as long as they conceivably could," said Lawrence Cohn, analyst with Merrill Lynch, Pierce, Fenner and Smith.

Cohn said the longer BankAmerica waits, the longer they have to show the market an improved financial outlook.

Although BankAmerica has yet to specify the types of equities it would offer, most analysts believed a convertible preferred stock would encompass at least part of it.

Such an offering at a depressed stock price would mean a lower conversion price and more dilution to BankAmerica stock holders, noted Daniel Williams, analyst with Sutro Group.

Several analysts said that while they believe the Brazilian debt problem will continue to hang over the banking industry through the quarter, the initial shock reaction is likely to ease over the coming weeks.

Nevertheless, BankAmerica, which holds about 2.70 billion dlr in Brazilian loans, stands to lose 15-20 mln dlr if the interest rate is reduced on the debt, and as much as 200 mln dlr if Brazil pays no interest for a year, said Joseph Arsenio, analyst with Birr, Wilson and Co.

He noted, however, that any potential losses would not show up in the current quarter.

With other major banks standing to lose even more than BankAmerica if Brazil fails to service its debt, the analysts said they expect the debt will be restructured, similar to way Mexico's debt was, minimizing losses to the creditor banks.

Reuter

</BODY></TEXT>

</REUTERS>

Appendix H – Answer Biographical Questions

A sample biography from Biography.com

Source: <http://www.biography.com/people/cleveland-abbe-9173762>

FOLLOW BIO: FACEBOOK • TWITTER • MOBILE • EMAIL UPDATES

Sign in with Facebook | Sign in | Register

bio. TRUE STORY: PEOPLE • TV • VIDEO • BIO NOW • ON THIS DAY • SHOP

Alvar Aalto

CELEBRITY NIGHTMARES DECODED NEW EPISODES SATURDAYS 10/9C

SEE WHAT'S ON TONIGHT!

商业新闻

新的网上兼职工作，每月可以赚 41000元以上。 [现在申请](#)

ADVERTISEMENT

Cleveland Abbe. biography

+ share

Home • People • Cleveland Abbe

Synopsis

[Print](#) [Cite This](#)

Cleveland Abbe inaugurated a public weather service that served as a model for the national weather service, which was organized shortly thereafter as a branch of the U.S. Army Signal Service. In 1871 he was appointed chief meteorologist of the branch, which in 1891 was reorganized as the U.S. Weather Bureau (later the National Weather Service), and he served in that capacity more than 45 years.

QUICK FACTS

NAME: Cleveland Abbe
OCCUPATION: Inventor, Astronomer, Meteorologist
BIRTH DATE: December 03, 1838
DEATH DATE: October 28, 1916
EDUCATION: Free Academy
[more about Cleveland](#)

BEST KNOWN FOR

Meteorologist Cleveland Abbe inaugurated a public weather service that served as a model for the U.S. Weather Bureau, later called the National Weather Service.

Profile

(born Dec. 3, 1838, New York, N.Y., U.S.—died Oct. 28, 1916, Chevy Chase, Md.) U.S. meteorologist. He was trained as an astronomer and appointed director of the Cincinnati Observatory in 1868. His interest turned to meteorology, and he inaugurated a public weather service that served as a model for the national weather service, which was organized shortly thereafter as a branch of the (U.S. Army) Signal Service. In 1871 he was appointed chief meteorologist of the branch, which in 1891 was reorganized under civilian control as the U.S. Weather Bureau (later the National Weather Service), and he served in that capacity more than 45 years.

Copyright © 1994-2011 Encyclopædia Britannica, Inc. For more information visit Britannica.com

CONTENTS

[Synopsis](#)
[Profile](#)

ADVERTISMENT

430元/小时 兼职
失业母亲通过网络工作赚43400元/月！
[了解她是怎么做到的](#)

24岁，每周在网上赚11100元
[了解她是怎么做到的。](#)
[阅读更多](#)

YOUR CONNECTIONS

Sign in with Facebook to see how you and your friends are connected to famous icons.

Sign in with Facebook

PROFILE CONNECTIONS

Ben Gazzara was also born in New York, New York

Len Lesser was also born December 3

A sample biography from Wikipedia

Source: [http://en.wikipedia.org/wiki/Charles_Clark_\(governor\)](http://en.wikipedia.org/wiki/Charles_Clark_(governor))

Wikipedia: Charles Clark (governor)

[Top](#)

[Home](#) > [Library](#) > [Miscellaneous](#) > [Wikipedia](#)

Charles Clark (May 24, 1811 – December 18, 1877) was a [Mississippi Democratic political figure](#), as well as a [major general](#) in the [Confederate States Army](#) during the [American Civil War](#).

Contents [\[hide\]](#)

- [1 Early life and career](#)
- [2 Civil War](#)
- [3 See also](#)
- [4 References](#)
- [5 Notes](#)
- [6 External links](#)

Early life and career

Clark was born in [Cincinnati, Ohio](#), in 1811. He subsequently moved to Mississippi.

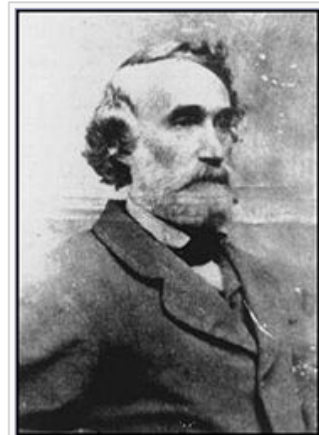
In the late 1830s and early 1840s, Clark, a [lawyer](#), represented a settler in a dispute with some [Choctaw](#) Native Americans over land in the [Mississippi Delta](#). The dispute led to a series of lawsuits before the [Mississippi Supreme Court](#). The settler ultimately prevailed, and gave Clark a large tract of land between [Beulah, Mississippi](#) and the [Mississippi River](#) as his legal fee. In the late 1840s, Clark formed a [plantation](#) on the land, naming it Doe-Roe, [pseudonyms](#) commonly used in the legal profession to represent unnamed or unknown litigants (e.g., [John Doe](#), [Roe v. Wade](#)). The state of literacy being what it was at the time, however, the plantation came to be known by its [phonic](#) representation, Doro. According to archives at [Delta State University](#), "The plantation grew to over 5,000 acres (20 km²) and became the most prosperous in the region, operating until 1913. It was prominent in the social, political and economic affairs of Bolivar County."^[1]

Civil War

Following the [secession](#) of Mississippi in early 1861, Clark was appointed as a [brigadier general](#) in the Mississippi 1st Corps, a state [militia](#) organization that later entered the Confederate Army. He commanded the brigade at engagements in [Kentucky](#) and then a [division](#) under [Leonidas Polk](#) at the [Battle of Shiloh](#). He was promoted to the rank of major general of Mississippi State Troops in 1863. Clark led a division at the [Battle of Baton Rouge](#), where he was severely wounded and captured. He spent time as a [prisoner of war](#) before being released.

On November 16, 1863, Clark was inaugurated as [governor of Mississippi](#) under [Confederate](#) auspices. He served in this capacity until June 13, 1865, when he was forcibly removed from office by occupation forces of the [United States Army](#) and replaced by [William L. Sharkey](#), a respected [judge](#) and staunch Unionist who had been in total opposition to [secession](#). Clark was imprisoned at [Fort Pulaski](#) near [Savannah, Georgia](#).

He died in [Bolivar County, Mississippi](#), on December 18, 1877, and was buried at the family graveyard in that county.^[2]



Charles Clark

A sample biography from astrotruths.com

http://www.astrotruths.com/noste/bio_nostradamus.php



- Home
- Seduce Lover by Astrology
- Detail Horoscope
- Nostradamus Topics
- Vedic Hindu Astrology
- Chinese Astrology
- Numerology
- Tantra Secret
- Mystic Mantras
- Tarot Card Reading
- Baby Names
- Contact Us

[Member login](#) | [New User Register](#)

Biography of Nostaedamus

On December 14, 1503 in St. Remi, France, Michel de Nostredame was born. The first son of Jewish parents, forced by the Inquisition to convert to Catholicism, would become a skilled physician but would gain renown during his lifetime and beyond as a seer of the future.



Growing up he spent much of his time learning languages, math, astronomy, and astrology from his grandfather, Jean. Later he attended the University at Avignon where he studied liberal arts. Afterwards, he graduated from the medical school at the University of Montpellier and began a private practice where he succeeded at treating plague victims in Montpellier and the surrounding areas.

Around 1534 he married and began a family. Tragically, the plague which he had been so successful in treating previously took the lives of his wife and two children. (The names of his wife and children are not known)

Distraught and pursued by the Inquisition, Nostradamus packed his bags and traveled throughout Italy and France for the next six years.

He eventually settled down in the town of Salon, France in 1554 where he married his second wife, Anne Ponsart Gemelle, with whom he raised six children - three boys and three girls.

It was during this time that he began his career as a prophet. In 1555, at the age of 52, he wrote his first collection of Centuries - a set of 100 quatrains. Over the next several years he would complete a total of 10 Centuries.



In 1564 Nostradamus was appointed Royal Physician to King Charles IX.

On July 1, 1566 Nostradamus offered his final prediction to his priest. In response to the priest's farewell of "Until tomorrow," Nostradamus is said to have answered: "You will not find me alive at sunrise." Nostradamus died that night. (Back to the home) (Tell your friends about this page)

[the prophecies of nostradamus](#)

Nostradamus 2012 Prediction Prophec University of Metaphysical Sciences

UMSonline.org/2012NostradamusProph

[Free Horoscopes Forecast](#)

Future, Love, Fortune Forecast and Lucky Numbers revealed. All Free

sara-freder.com/Renowned-Astrologer

[Passover 2010 Resorts](#)

New Passover Vacations Directory Passover 2010 in the USA, Europe...

www.Pesach-holidays.com

[Conspiracy Theories](#)

Misinformation & conspiracy caused by fear and ignoring the reality

www.america.gov



Ads by Google

H.2 AN EXAMPLE BIOGRAPHICAL FEATURE SET

Feature
DATE in LOCATION
LOCATION, ORGANIZATION
LOCATION, where
ORGANIZATION in YEAR
PERSON (YEAR-YEAR
PROFESSION,
SINGLE_PERSON_PRONOUN
PROFESSION, born
PROFESSION and PROFESSION
SINGLE_PERSON_PRONOUN became
SINGLE_PERSON_PRONOUN became a
SINGLE_PERSON_PRONOUN began
SINGLE_PERSON_PRONOUN career
SINGLE_PERSON_PRONOUN founded
SINGLE_PERSON_PRONOUN joined
SINGLE_PERSON_PRONOUN later
SINGLE_PERSON_PRONOUN made
SINGLE_PERSON_PRONOUN published
SINGLE_PERSON_PRONOUN was
SINGLE_PERSON_PRONOUN was appointed
SINGLE_PERSON_PRONOUN was elected
SINGLE_PERSON_PRONOUN work
SINGLE_PERSON_PRONOUN wrote

Feature
a PROFESSION
a PROFESSION of
a member of the
after
age
appointed
as PROFESSION
as a PROFESSION
at ORGANIZATION
became PROFESSION
born in LOCATION
died
discovered
during the
educated
elected
elected to
founded the
from YEAR
from YEAR to
in LOCATION
in YEAR
married

TABLE H-1. EXAMPLE BIOGRAPHICAL FEATURE SET