Unsupervised Relation Learning for Event-Focused Question-Answering and Domain Modelling

Elena Filatova

Submitted in partial fulfillment of the
requirements for the degree
of Doctor of Philosophy
in the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2008

ABSTRACT

Unsupervised Relation Learning for Event-Focused Question-Answering and Domain Modelling

Elena Filatova

> It is a very sad thing that
> nowadays there is so little
> useless information.
>
> Oscar Wilde

In this thesis, we investigate the problem of identifying, within a text, relations that capture information important for event-focused document collections. The presented solutions work with events of various granularity and we show how to use these relations to improve the performance of a number of natural language processing applications.

For a set of related event-focused documents, we introduce a notion of a *shallow semantic network* based on the relations between the important elements discovered in these documents. This shallow semantic network captures the most important relations among the objects, people, and other elements that are involved in the events described in the input document collection. We present experimental evidence that such a relation-based representation of event-focused documents is superior to techniques that rely on term frequencies for the task of information selection.

For a set of document collections describing similar events within the same domain, we design and implement a completely automatic, data-driven procedure for inducing domain templates. These domain templates capture facts that are important for most domain instances.

We then devise a procedure for identifying commonalities across different subdomains. We experiment with a special case of a domain, a *biography* domain, and identify commonalities across activities used for descriptions of people belonging to different occupations.

We also propose a methodology for creating domain hierarchies.

We apply our methods for identifying relations to the question-answering task. We design and implement a two-pronged approach for answering open-ended event-related questions. The first approach relies on automatically created domain templates and is used when the event mentioned can be identified as one of a particular class of events (e.g., earthquakes, presidential elections). The second, complementary approach is based on a shallow semantic network, which we extract from the documents relevant to the question.

We also suggest a formal model for efficient information packaging that is based on mapping the information selection task onto the set cover complexity problem. Using this mapping, we outline and implement information selection algorithms that are provably optimal polynomial-time approximations for information selection tasks that have a limit on the output size.

Contents

1 Introduction 1

1.1 Thesis Contributions 2

1.2 Thesis Organization 4

2 Event Descriptions in Text 7

2.1 Events in Linguistics Literature 8

2.1.1 Narrative Theory 8

2.1.2 Verb classes 8

2.2 TimeML language specification and TimeBank corpus 9

2.3 Events as Text Snippets 11

2.4 Events within the Information Extraction Task 12

2.5 Events in Information Retrieval (TDT) 14

2.6 Conclusions 15

3 Semantic Network and Information Selection 17

3.1 Related Work 19

3.2 A Study of Event Annotation 22

3.2.1 Length of Marked Text Passages 24

3.2.2 Text Features in Marked Passages 25

3.3 Identifying Features for Event Description 27

3.4 Shallow Semantic Network 28

3.5 System Evaluation 33

3.5.1 Methodology 33

3.5.2 Results 35

3.6 Atomic Relations and Information Extraction 36

3.6.1 History of Information Extraction (IE) 36

3.6.2 Overview of IE Systems 37

3.6.2.1 Extraction Systems Dealing with Pre-Specified Relation Types (Classic IE Systems) 38

3.6.2.2 Extraction Systems Dealing with Extraction of Relations of All Types from the Input Corpus 41

3.6.3 Comparison of Atomic Relations and IE Systems 43

3.7 Formal Model for Information Packaging and Selection 45

3.7.1 Full Correspondence 46

3.7.2 Partial Correspondence between Textual and Conceptual Units . . . 47

3.7.3 Length and Textual Constraints 47

3.7.4 Handling Redundancy in Summarization 49

3.7.5 Applying the Model 50

3.8 Experiments 52

3.8.1 Results: Static Greedy Algorithm 54

3.8.2 Results: Adaptive Greedy Algorithm 58

3.8.3 Results: Modified Adaptive Greedy Algorithm 63

3.9 Results Discussion 64

3.9.1 Conclusion 67

4 Domain Modelling **69**

4.1 Related Work 71

4.2 Our Approach to Template Induction 75

4.3 Data Description 76

4.3.1 Training Data 76

4.3.2 Test Data 76

4.4 Inducing Templates 77

4.5 Discussion 81
4.6 Evaluation 82
4.6.1 Stage 1. Information Included into Templates: Interannotator Agreement 83
4.6.2 Stage 2. Quality of the Automatically Created Templates 86
4.7 IE and Automatic Domain Template Induction 90
4.8 Conclusions 91

5 Columbia Event QA System 93
5.1 Related Work 95
5.1.1 Factoid Questions 97
5.1.2 Questions Requiring Long Answers 98
5.1.3 The Specifics of Event-Related Questions 100
5.2 Columbia QA System for Answering Open-Ended Event-Related Questions 101
5.2.1 Pipeline Overview 102
5.2.2 Event-Related Questions within GALE 104
5.2.3 Evaluation Set-Up 106
5.3 Answering *Template 1* Questions (Prong 1) 108
5.3.1 System Version 1 109
5.3.1.1 System Version 1: Information Retrieval Stage 109
5.3.1.2 System Version 1: Sentence Selection Stage 110
5.3.1.3 System Version 1: Evaluation 112
5.3.1.4 System Version 1: Error Analysis 114
5.3.2 System Version 2 116
5.3.2.1 System version 2: Information Retrieval Stage 116
5.3.2.2 System Version 2: Sentence Selection Stage 117
5.3.2.3 System version 2: Evaluation 117
5.3.3 System Version 3 119
5.3.3.1 System Version 3: Information Retrieval Stage 119

5.3.3.2 System Version 3: Sentence Selection Stage 120
5.3.3.3 System Version 3: Evaluation 121
5.3.4 System Limitations 123
5.4 Answering *Template 8* Questions (Prong 2) 123
5.4.1 Domain Template Generation Stage 124
5.4.2 Information Retrieval Stage 125
5.4.3 Sentence Selection Stage 126
5.4.4 Evaluation 127
5.5 Conclusions and Future Work 130

6 Biography Domain 133

6.1 Related Work 134
6.2 Unsupervised Learning of Activities Typical for Different Occupations . . . 138
6.2.1 Data 138
6.2.2 Extracting Occupation-Specific Activities 140
6.2.3 Generalized Atomic Relations 142
6.3 Getting Occupation-Related Activities 143
6.4 Classification 146
6.4.1 SVM Classification Using Only Verbs 147
6.4.2 SVM Classification Using Atomic Relations 150
6.4.3 Other Types of Classification 151
6.4.4 Using Classification Extracted Features 152
6.5 Clustering 153
6.5.1 CLUTO CLustering Tool 155
6.5.2 Identifying the Number of Clusters 156
6.6 Conclusions and Future Work 160

7 Contributions, Limitations and Future Work 163

7.1 Contributions 163

7.2 Limitations and Future Work 165

Bibliography **167**

Appendices **189**

A MUC-4 **190**

A.1 General Template for Terrorist Attacks Domain 191

A.2 Document TST4-MUC4-0004 192

A.3 Key for TST4-MUC4-0004 193

A.4 Document TST4-MUC4-0006 194

A.5 Key for TST4-MUC4-0006 195

B Answers to Template 1 Questions **196**

B.1 Answers to GALE Year 1 Go/No-Go Evaluation Template 1 questions, Version 1 (Year 1 Go/No-Go system) 196

B.2 Answers to GALE Year 1 Go/No-Go Evaluation Template 1 questions, Version 2 (with two-step IR) 214

B.3 Answers to GALE Year 1 Go/No-Go Evaluation Template 1 questions, Version 3 (with two-step IR and information packageing procedure) 229

C Answers to Template 8 Questions **248**

C.1 Answers Created Using the Automatically Induced Domain Template . . . 248

C.2 Answers Created Using a Shallow Semantic Network 251

List of Figures

2.1 Events pipeline. The portion of this pipeline that was implemented within the Columbia QA system for answering event-related open-ended questions is described in detail in Chapter 5. 16

3.1 Official description of *China Airlines crash* topic. 29

3.2 Screen shot of the system output. 34

3.3 Example of an FP-structure. 39

3.4 Average ROUGE-1 and ROUGE-L scores over 30 DUC 2001 documents (static algorithm, exactly 50 words). 56

3.5 Average ROUGE-1 and ROUGE-L scores over 30 DUC 2001 documents (static algorithm, exactly 100 words). 57

3.6 Average ROUGE-1 and ROUGE-L scores over 30 DUC 2001 documents (static algorithm, exactly 200 words). 57

3.7 Average ROUGE-1 and ROUGE-L scores over 30 DUC 2001 documents (static algorithm, exactly 400 words). 58

3.8 ROUGE-1 scores over 30 DUC 2001 documents (static algorithm, exactly 50 words). 58

3.9 ROUGE-1 scores over 30 DUC 2001 documents (static algorithm, exactly 100 words). 59

3.10 ROUGE-1 scores over 30 DUC 2001 documents (static algorithm, exactly 200 words). 59

3.11 ROUGE-1 scores over 30 DUC 2001 documents (static algorithm, exactly 400 words). 60

3.12 Average ROUGE-1 and ROUGE-L scores over 30 DUC 2001 documents (adaptive algorithm, exactly 50 words). 60

3.13 Average ROUGE-1 and ROUGE-L scores over 30 DUC 2001 documents (adaptive algorithm, exactly 100 words). 61

3.14 Average ROUGE-1 and ROUGE-L scores over 30 DUC 2001 documents (adaptive algorithm, exactly 200 words). 61

3.15 Average ROUGE-1 and ROUGE-L scores over 30 DUC 2001 documents (adaptive algorithm, exactly 400 words). 62

3.16 ROUGE-1 scores over 30 DUC 2001 documents (adaptive algorithm, exactly 50 words). 62

3.17 ROUGE-1 scores over 30 DUC 2001 documents (adaptive algorithm, exactly 100 words). 63

3.18 ROUGE-1 scores over 30 DUC 2001 documents (adaptive algorithm, exactly 200 words). 63

3.19 ROUGE-1 scores over 30 DUC 2001 documents (adaptive algorithm, exactly 400 words). 64

3.20 Average ROUGE-1 and ROUGE-L scores over 30 DUC 2001 documents (modified adaptive algorithm, exactly 50 words). 65

3.21 Average ROUGE-1 and ROUGE-L scores over 30 DUC 2001 documents (modified adaptive algorithm, exactly 100 words). 65

3.22 Average ROUGE-1 and ROUGE-L scores over 30 DUC 2001 documents (modified adaptive algorithm, exactly 200 words). 66

3.23 Average ROUGE-1 and ROUGE-L scores over 30 DUC 2001 documents (modified adaptive algorithm, exactly 400 words). 66

3.24 ROUGE-1 scores over 30 DUC 2001 documents (modified adaptive algorithm, exactly 50 words). 67

3.25 ROUGE-1 scores over 30 DUC 2001 documents (modified adaptive algorithm, exactly 100 words). 67

3.26 ROUGE-1 scores over 30 DUC 2001 documents (modified adaptive algorithm, exactly 200 words). 68

3.27 ROUGE-1 scores over 30 DUC 2001 documents (modified adaptive algorithm, exactly 400 words). 68

4.1 Restaurant visit script (S - customer). 71

4.2 Examples of manually created case frames for the *terrorist attack* domain. . 73

4.3 PropBank annotation guidelines for the verb *to kill*. 81

4.4 Evaluation results. 88

5.1 Pipeline for Columbia event-related open-ended QA system. 102

5.2 Official guidelines for Template 1 questions. 105

5.3 Official guidelines for Template 8 questions. 106

5.4 Precision scores for the seven Template 1 questions used in Year 1 GALE Go/No-Go formal distillation evaluation (system version 1). 113

5.5 Recall scores for the seven Template 1 questions used in Year 1 GALE Go/No-Go formal distillation evaluation (system version 1). 114

5.6 F-value scores for the seven Template 1 questions used in Year 1 GALE Go/No-Go formal distillation evaluation (system version 1). 115

5.7 Precision scores for the seven Template 1 questions used in Year 1 GALE Go/No-Go formal distillation evaluation (system version 2). 118

5.8 Recall scores for the seven Template 1 questions used in Year 1 GALE Go/No-Go formal distillation evaluation (system version 2). 119

5.9 F-value scores for the seven Template 1 questions used in Year 1 GALE Go/No-Go formal distillation evaluation (system version 2). 120

5.10 Precision scores for the seven Template 1 questions used in Year 1 GALE Go/No-Go formal distillation evaluation (system version 3). 121

5.11 Recall scores for the seven Template 1 questions used in Year 1 GALE Go/No-Go formal distillation evaluation (system version 3). 122

5.12 F-value scores for the seven Template 1 questions used in Year 1 GALE Go/No-Go formal distillation evaluation (system version 3). 123

5.13 Precision scores for the seven Template 1 questions used in Year 1 GALE Go/No-Go formal distillation evaluation (system version 3, BBN, and IBM systems). 124

5.14 Recall scores for the seven Template 1 questions used in Year 1 GALE Go/No-Go formal distillation evaluation (system version 3, BBN, and IBM systems). 125

5.15 F-value scores for the seven Template 1 questions used in Year 1 GALE Go/No-Go formal distillation evaluation (system version 3, BBN, and IBM systems). 126

5.16 The Indri query for the Template 8 question DESCRIBE THE PROSECUTION OF [*Saddam Hussein*] FOR [*crimes against humanity*]. 127

5.17 Precision scores for the five Template 8 questions used in Year 1 GALE Go/No-Go formal distillation evaluation. 128

5.18 Recall scores for the five Template 8 questions used in Year 1 GALE Go/No-Go formal distillation evaluation. 129

5.19 F-value scores for the five Template 8 questions used in Year 1 GALE Go/No-Go formal distillation evaluation. 130

5.20 Number of answer sentences produced for the five Template 8 questions used in Year 1 GALE Go/No-Go formal distillation evaluation. 131

6.1 A segment of the Semantic Web pertaining to Yo-Yo Ma. 136

6.2 Bipartite graph for a set of generalized atomic relations corresponding to *explorers* occupation 144

6.3 Grouping people into ten clusters. 157

6.4 Grouping people into two clusters. 159

6.5 Grouping people into eleven clusters. 159

A.1 MUC–4 template for *Terrorist Attacks* domain 191

A.2 Filled-in tempalte for document TST4-MUC4-0004 193

A.3 Filled-in tempalte for document TST4-MUC4-0006 195

List of Tables

2.1 IE template for a terrorist attack event. 12

2.2 ACE *Life* event subtypes. 13

3.1 Length of text snippets covering event descriptions. 24

3.2 Top 10 named entity pairs for the *China Airlines crash* topic. 30

3.3 Top connectors for three of the relations in Table 3.2. 31

3.4 Final event output for the relations of Table 3.2. 31

3.5 Top connectors across the entire *China Airlines crash* topic. 32

3.6 Pairs which are and are not specific for the *China Airlines crash* topic. . . . 32

3.7 Event extracted for the noun "passengers" from the *China Airlines crash* topic. 33

3.8 Evaluation scores for our system. Importance and label quality measured only on extracted relations of reasonable quality (with link quality score above 0.5, 75% of the total extracted events). 35

3.9 T-test values for all the text unit selection modes, and summary lengths. . . 55

4.1 Sample subtrees for the *terrorist attack* domain. 79

4.2 Creating gold standard. Jaccard metric values for interannotator agreement. 85

5.1 ACE *Life* event subtypes. 101

6.1 Answers for "When did Leonardo da Vinci paint the Mona Lisa?" 136

6.2 A sample of atomic relations extracted for the collection of documents about Christopher Columbus 141

6.3 A sample of atomic relations extracted for two representatives of the explorer occupation 142
6.4 Top ten activities for three occupations: dancers, physicists and singers . . 145
6.5 Top ten activities common for all the eleven occupations. 147
6.6 Performance of different classification methods. 148
6.7 Occupation-specific generalized atomic relations (note that NAME stands for NAME/PERSON). 154
6.8 Generalize atomic events typical for any biography 154
6.9 Most descriptive generalized atomic relations (biographical activities) for the clusters presented in Figure 6.3. 162

Acknowledgments

There are a number of people to whom I am indebted for encouraging and helping me throughout my academic journey. First and foremost, I would like to thank my advisor, Kathy McKeown, who believed in me even when I stopped believing in me myself. Kathy's understanding, encouragement, and support pushed me forward and were the driving forces behind my research. But Kathy's help extended much beyond research advice. She is a great and very warm person who cares not only about the success of her students in the academic world but also about their happiness in every-day life. The NLP group she leads in Columbia has a great camaraderie feeling and is more than a group of fellow researchers but also friends and supporters. Owen, Mona, Nizar, Michel, Davids, Carl, Kapil, Kristen, Barry, Smara, Sashas, you are the best.

I am thankful to Eduard Hovy. He taught me the basics of conducting research and became not only my academic mentor but also a patient and kind friend who helped me a lot during the years I spent in Los Angeles. Even after I left USC he stayed very much interested in the progress I was making and was always willing to share his wisdom. I also wish to thank all the members of the ISI NLP group for being supportive and understanding and for helping me starting life far away from home. If not for Laurie, Kevin, Daniel, Uli, Ulf, Kristina, Katya, Aram and everyone else's help I would have left Los Angeles and gone back home in two weeks after my arrival.

There are many other people without whom this dissertation would not have happened. I apologize in advance to everyone whom I do not mention here due to lack of space. Believe me, I remember and thank each and everyone of you. I indebted to Vasileios Hatzivassiloglou for bringing me to Columbia. I am honored to have met and worked with John Prager and Julia Hirschberg. I am grateful to my professors from Moscow State University. I am thankful to my friends all over the world. And it goes without saying that I would not be where I am right now without the main people in my life: my family.

Chapter 1

Introduction

We often refer to our era as the "information age." Information has become a valuable commodity and a priceless asset. Plenty of information exists in text form. Some of this textual information is structured (e.g., databases), but a great deal of information is stored as unstructured text (e.g., books, news articles, reports). For example, many people use Google on a daily basis to retrieve sets of documents containing information of interest. Although the quality of modern search engines is very high, a user still must read and analyze many retrieved documents to locate the information that answers the submitted query, especially when the answer is a combination of information from multiple web-pages. Obviously, it is desirable to automate the process of sifting through the retrieved documents, so that the user can be presented with an answer in the form of concise text, rather than with a list of documents corresponding to the query.

In general, the goal of Question Answering (QA) systems is: given a collection of documents (such as the World Wide Web or a local collection), retrieve answers to questions posed in natural language. Our research within the Question Answering task focuses on the problem of selecting information that should be included in answers to *questions about events* (e.g., "What happened during the presidential election in country X?" or "Describe the earthquake that happened in place Y."). Answering event-related questions is a challenging task as users often do not provide precise constraints on the information they expect in the answer. For example, in the above question about a presidential election, there is no explicit mention of vote distribution, campaigning, or candidate names, but it is highly likely that the user expects information about these aspects to be included in the answer. Thus, inducing from an event description its constituent parts and connections among these constituent parts (e.g., participants, locations, dates) is a crucial step in answering event-related questions. Moreover, if it is possible to identify what relations are important and descriptive for the domain to which the event in question belongs, these relations can be used for an answer sentence selection procedure targeted towards covering information important for a particular domain. Often, answers created using domain descriptive relations have very high precision. For example, it is clear that for the *presidential election* domain, information about who won the election, what the vote distribution was, and on what day the election was held, are expected to be included in the description of any presidential election irrespective of the election procedure used in the country. We call a set of relations that are generally important for a domain a *domain template.* However, deciding what relations are important for a particular domain is a difficult task that is one of the central tasks of this dissertation.

1.1 Thesis Contributions

In this thesis, we investigate the problem of how to identify, within a text, relations that capture information important for document collections of various granularities and how to use these relations to improve the performance of a number of natural language processing applications. The key contributions of this thesis are as follows:

- We introduce the notion of *shallow semantic network.* This network consists of *atomic relations* that connect the entities described in a collection of documents to each other. The central elements in the shallow semantic network are verbs and action nouns. We show that for the task of information selection from event-focused document collections, a relation-based representation of documents is superior to techniques that rely on term frequencies. Our results show that a shallow semantic network constructed out of atomic relations can be successfully used to capture succinctly the essence of the *events* described in a collection of text documents.

- We design and implement a novel, completely automatic data-driven procedure for *domain template* induction. Domain templates contain facts that are important for the respective domains. Summarization and QA systems have successfully used domain templates for targeted information selection. Currently, domain templates are typically created manually and thus, are time-consuming to create and are not portable across domains. Our approach uses unsupervised learning over descriptions of similar events and requires as input only a set of document collections corresponding to several instances of a particular domain. Given this set of document collections where each collection describes an event of the pre-defined type (domain), the domain template induction procedure performs cross-comparison of the information covered by these instances and deduces those information facts that are most descriptive for the domain under consideration. For example, comparing document sets describing different instances of presidential elections that took place in different countries on different dates we can learn that all these document sets contain information about campaigning and voting. Thus, these two information facts are good candidates to be included in the template for the *presidential election* domain. Our experiments show that the automatically created domain templates capture information facts that correspond well to human expectations on what information is important for a particular domain.

- We study a special case of a domain, the ***biography*** domain. Using this domain, we devise an approach for identifying commonalities across a set of different domains. We show how random walk theory can be used to identify information relevant to all the domains from the initial domain set, information relevant to a subset of the initial domain set, as well as the information relevant to a particular domain from the domain set. For the biography domain, this three-tier information distribution corresponds to three levels of activities: general biographic, occupation-related and person-specific. By building on these ideas we also show how to generate domain hierarchies.

- We introduce a two-pronged approach for answering event-related open-ended questions. This approach utilizes two novel techniques presented in this dissertation: shallow semantic network extraction and domain template induction. When the event in question cannot be classified as one of an *a priori* list of event-types, we learn the important event constituents on-the-fly from document collection corresponding to the question and use a shallow semantic network representation to identify those sentences that should be included in the answer. When the event in question can be recognized as one of a general class of events we use automatically acquired domain templates to guide the selection of answer information (e.g., relevant subevents, causes, effects). We also design and implement a two-step answer selection procedure that interleaves information retrieval and answer generation stages within a question-answering system. In the first step, only highly relevant documents are analyzed and a shallow semantic network is constructed for this small set of documents; in the second step, this shallow semantic network is propagated to a larger set of documents that have more information about the important event elements identified at the first step. We show the advantage of interleaving document retrieval and information selection steps.

- We introduce a formal model of efficient information packaging for information selection tasks that have limits on the output size. The main goal of such information selection tasks is to cover the maximal amount of information within a limited space. This model is based on mapping the information selection task onto the set cover complexity problem. Using this mapping, we outline and implement information selection algorithms that are provably optimal polynomial-time approximations for information selection tasks that have a limit on the output size.

1.2 Thesis Organization

In Chapter 2, we describe various approaches that are currently used by Natural Language Processing (NLP) systems for identifying events. We show, that although the notion of an *event* seems to be very intuitive and straight-forward, the variety of constructions that are identified in text and considered to correspond to event descriptions vary greatly in both structure and length. This chapter surveys multiple approaches for defining events in the field of NLP. We show that the existing various methods of event identification cannot and should not compete against each other. Instead, the combination of different event identification systems can lead towards the creation of a very powerful end-to-end system which, given a query, would retrieve appropriate documents, identify pieces of text containing important information, assign time-stamps to these text pieces and finally, output a chronologically correct text (e.g., summary or an answer to a question). We describe a general framework within which various approaches for event identification can effectively collaborate.

In Chapter 3, we introduce the notion of shallow semantic network. This network connects objects that are considered to be important constituent parts of the events described in text. In this chapter we also introduce a formal model for information packaging based on the set cover complexity problem. We show that, combined, the shallow semantic network and the information selection model can be successfully used for event-focused summarization. The key contribution of this chapter is two-fold: First, we introduce a novel approach for event-focused document collection representation (shallow semantic network). This shallow semantic network captures the most important relations among the objects, people, and notions that are involved in the events described in the input document collection. Second, we describe how the information packaging problem can be mapped onto a well-defined complexity problem. This mapping demonstrates the hardness of the information selection process and suggests a provably optimal approximate solution.

A set of document collections describing similar events (for example, different cases of presidential elections) can be used for further development of the shallow semantic network technique. In Chapter 4, we describe a data-driven procedure for automatic domain template induction. Previously, domain templates were created manually; this procedure is time-consuming, and the created templates are not portable across domains. The key contribution of this chapter is a data-driven, completely automatic procedure for domain template induction based on cross-comparison of a set of domain instances. We evaluate the quality of the induced domain templates and show that the information captured according to these templates corresponds well to the user expectations on what information is important for the domain.

In Chapter 5, we describe the problem of QA in general and open-ended event-related QA in particular. We describe a two-pronged approach for answering open-ended event-related questions. This two-pronged approach utilizes the novel techniques presented in Chapters 3 and 4. We present the evaluation results of the QA system for answering event-related questions using a shallow semantic network and an automatically induced domain template. We experimentally show how a combination of information retrieval and general summarization techniques can be repurposed to better suit the task of open-ended QA.

In Chapter 6, we explore the synergy of shallow semantic networks and domain modelling for a biography domain. We show how a generalized version of a shallow semantic network can be used to create a domain template capturing information about a particular occupation or a domain template containing general biographical facts. We present a novel task of classifying people according to their occupations based on the occupation-related domain templates. We show how shallow semantic networks that correspond to people descriptions can be used not only as classification features but as clustering features as well. We demonstrate that, using a different number of clusters, it is possible to create a hierarchy of occupations/domains whose nodes correspond to occupation names and contain information about representatives of these occupations as well as occupation-related activities typical for these occupations.

In Chapter 7, we summarize the contributions of this dissertation, discuss the limitations of the suggested algorithms, and discuss the directions for future research.

Chapter 2

Event Descriptions in Text

> Life is so constructed, that the
> event does not, cannot, will not,
> match the expectation.
>
> Charlotte Bronte

The notion of *event* has been widely used in the Natural Language Processing (NLP) literature as well as in other related fields (e.g., Information Retrieval), although with significant variance in what exactly an event is. Often researchers do not give any formal definition of what an event is, but rather use an intuitive definition, similar to the one presented by WordNet[1] that defines an event broadly as "something that happens at a given place and time". Unfortunately, this definition is generic and a variety of different text constructions satisfy it. In this Chapter we present a variety of approaches for event realization in text. The text constructions that correspond to event descriptions differ in structure, length, and the amount and type of information covered by them. The wide variation in what different researchers mean by the term *event* makes communication among researchers difficult. In this chapter, we present an analysis of the literature on event identification and extraction. We cover existing approaches for defining events in NLP and other related fields, and describe the text constructions that correspond to these events. We show that these different types of events are identified in different text levels (from a

[1] `http://www.cogsci.princeton.edu/~wn/`

single word to a document collection). We also suggest a system architecture where the existing event types can be used synergistically to provide valuable information about the input text.

2.1 Events in Linguistics Literature

2.1.1 Narrative Theory

Event is an important concept for the *Theory of Action*. Among the questions central for the Theory of Action listed by Davidson (2001) are: "What events in the life of a person reveal agency; What are his deeds and his doings in contrast to mere happenings in history; what is the mark that distinguishes his actions?"

The core property of the *Narrative theory* is representing events or changes of states (Herman, Jahn & Ryan 2005) in comparison to property or general information descriptions (as shown below).

1. Property description:

   Water is H_2O.

2. Inductive generalization:

   Water freezes at 0 *degrees.*

3. Event:

   The temperature dropped to 0 *last week and the pond behind my house froze.*

The researchers working within the narrative theory constraints study events mainly from a theoretical point of view. The application of narrative theory for automatic analysis of text was developed within the research framework on verb classes.

2.1.2 Verb classes

The notion of an *event* and its semantic structure has been analyzed by several linguists, who have looked at semantic constraints in sentences to distinguish between events, extended events, and states; see for example (Chung & Timberlake 1985, Bach 1986, Pustejovsky 2000, Siegel 1999, Siegel & McKeown 2000). Often in such research, event analysis is centered on properties of the verb, and verbs are classified according to their relationships with event classes (Levin 1993).

Linguists who work on the underlying semantic structure of events and their realization in text propose a definition of an event involving telicity, time, and external world conditions; for example, Chung and Timberlake in (Chung & Timberlake 1985) say that "an event can be defined in terms of three components: a predicate; an interval of time on which the predicate occurs, which we call the event frame; and a situation or set of conditions under which the predicate occurs, which we call the event world." Siegel and McKeown (1999, 2000) have proposed automatic methods for classifying verbs according to whether they can signal events and processes (stativity and completeness).

The latest research of this type of events has led to a development of a new task in the Information Extraction field, namely assigning time-stamps to the events described in text.

2.2 TimeML language specification and TimeBank corpus

Pustejovsky in (Pustejovsky 2000) argues for a semantic theory of events that models persistence as well as change and is grounded in the notion of predicate opposition between objects and properties. He notes that "lexical semanticists must look outward from the verb to the sentence in order to characterize the effects of a verb's event structure; and logical semanticists must look inward from the sentence to the verb to represent semantic facts that depend on event-related properties of particular verbs".

To create machine tools capable of automatic identification of events expressed though verbs, nouns, adjectives, predicative clauses and prepositional phrases the TimeML specification language of new articles was created (Pustejovsky, Castaño, Ingria, Saurí, Gaizauskas, Setzer, Katz & Radev 2003). The TimeML specification language allows marking not only events but also assigning time-stamps to the marked events.

In TimeML an *event* is "a cover term for situations that happen or occur." "[..] predicates describing states or circumstances in which something obtains or holds true" are also considered to be events. TimeML events may be expressed by means of:[2]

1. tensed or untensed verbs

   A fresh flow of lava, gas and debris **erupted** *there Saturday.*

   Prime Minister Benjamin Netanyahu called the prime minister of the Netherlands **to thank** *him for thousands of gas masks his country has already contributed.*

2. nominalizations

   Israel will ask the United States to delay a military **strike** *against Iraq until the Jewish state is fully prepared for a possible Iraqi attack.*

3. adjectives

   A Philippine volcano, **dormant** *for six centuries, began exploding with searing gases, thick ash and deadly debris.*

4. predicative clauses

   "There is no reason why we would not **be prepared**, *" Mordechai told the Yediot Ahronot daily.*

5. prepositional phrases

   All 75 people **on board** *the Aeroflot Airbus died."*

There exist systems that identify events following the TimeML annotation guidelines automatically using rule-based (Saurí, Verhagen & Pustejovsky 2005) or machine learning (Bethard & Martin 2006) approaches.

The TimeML specification language was used to create the TimeBank (Pustejovsky, Hanks, Saurí, See, Gaizauskas, Setzer, Radev, Sundheim, Day, Ferro & Lazo 2003) corpus.

[2] The below event classification and examples are taken from the official TimeML Annotation Guidelines (Version 1.2) `http://www.cs.brandeis.edu/~jamesp/arda/time/timeMLdocs/annguide12wp.pdf` or `http://www.cs.brandeis.edu/~jamesp/arda/time/documentation/AnnotationGuidelineDraft2.html`.

The TimeBank corpus consists of 183 news articles that are annotated with temporal information. This information can be used for better understanding of temporal links between events and times. TimeBank has been used as a reference corpus to learn rules for annotating events with the respective duration time (Pan, Mulkar & Hobbs 2006*a*, Pan, Mulkar & Hobbs 2006*b*)

Other research projects that studied the problem of assigning time to events treated events as time spans rather than verbs, nouns or prepositional phrases. From a computational perspective, discourse analysis has relied on (often implicitly defined) events; for example, McCoy and Strube (McCoy & Strube 1999) investigated time intervals that can be assigned to events (atomic events occurred at a single point in time versus repeated atomic events, extended atomic events or states that occurred over a span of time) to generate pronouns. In this work simple clauses were taken as the text region for events. Filatova and Hovy (2001) also identified a simple clause as a text span for an event and studied the problem of assigning time-stamps to all the simple clauses into which the input text is divided.

Identifying a short text span as a description of an event has been studied not only for assigning time-stamps to events, but for other tasks as well (e.g., summarization).

2.3 Events as Text Snippets

The events described above were studied primarily as stand-alone events. There exist NLP systems, however, that study connections among events. For such tasks, event descriptions often correspond to spans of text. Daniel *et al* (2003) identify sub-events in a document collection and rank sentences according to their relevancy to these subevents.

Naughton *et al* (2006) use machine learning for identifying which sentences from news articles contain event descriptions and then cluster sentences containing descriptions of the same event.

There is no agreement on what is the optimal length of a text-span that covers an event. Usually, either a sentence or a simple clause are thought to be long enough to contain an event description and short enough to contain a description of only one event. The events that have been described so far have an explicit structure to encode relations among the constituent parts within an event.

2.4 Events within the Information Extraction Task

Another approach to defining events is used for Information Extraction (IE). "Information extraction is the automatic identification of selected types of entities, relations, or events in free text" (Grishman 2003). To identify an IE event is to extract fillers for a predefined event template. For example, an IE event coding information about a terrorist attack can be represented as a template shown in the left side of Table 2.1. This table also contains fillers for the terrorist attack event template extracted from the following text:

> 19 March - A bomb went off this morning near a power tower in San Salvador leaving a large part of the city without energy, but no casualties have been reported. According to unofficial sources, the bomb - allegedly detonated by urban guerrilla commandos - blew up a power tower in the northwestern part of San Salvador at 0650 (1250 GMT).

Template slots	Slot fillers
Incident	type bombing
Date	March 19
Location	El Salvador: San Salvador (city)
Perpetrator	urban guerilla commandos
Physical target	power tower
Human target	-
Effect on physical	target destroyed
Effect on human target	-
Instrument	bomb

Table 2.1: IE template for a terrorist attack event.

Life event subtype	Arguments
Be Born	Person, Time, Place
Die	Agent, Victim, Instrument, Time, Place
Injury	Agent, Victim, Instrument, Time, Place
Marry	Person, Time, Place
Divorce	Person, Time, Place

Table 2.2: ACE *Life* event subtypes.

IE events were identified by systems participating in the Message Understanding Conferences (MUC) (*Proceedings of the Fourth Message Understanding Conference (MUC–4)* 1992, *Proceedings of the Fifth Message Understanding Conference (MUC–5)* 1993, *Proceedings of the Sixth Message Understanding Conference (MUC–6)* 1995, Marsh & Perzanowski 1997). They used a variety of approaches for event identification: rule-based, lexicon-drive, machine learning, etc.

Recently, NIST has created a new evaluation effort for IE event identification, Automatic Content Extraction (ACE) (Doddington, Mitchell, Przybocki, Ramshaw, Strassel & Weischedel 2004). For the ACE task, the participating systems are supposed to identify several pre-defined semantic types of events (life, justice, transaction, conflict, etc.) together with the constituent parts corresponding to these events (agent, object, source, target, time, location, other). For example, Table 2.2 lists life events together with the arguments which should be extracted for these events.

Many of the ACE systems use the PropBank (Palmer, Gildea & Kingsbury 2005) semantic annotation. PropBank is a recent initiative which adds semantic tags to the PennTreeBank corpus.

The task of labeling important events mentioned in news articles has also been addressed by statisticians in social, economic and political sciences. As in the NLP IE field, such systems rely on a predefined set of event categories. Each of these event categories is defined by a set of verbs corresponding to it. There exist several such event typologies: IDEA (King & Lowe 2003), KEDS, CAMEO (Gerner, Schrodt, Francisco & Weddle 1994).

For example, in the IDEA event typology the *criticize or denounce* event category can be instantiated by the following verbs: *blame, find fault, censure, rebuke, "whistle blowing," vilify, defame, denigrate, condemn and name-calling.* For each of the verbs, also source (S) and target (T) agents are identified, if possible (King & Lowe 2003).

2.5 Events in Information Retrieval (TDT)

We have described approaches where the length of an event description in text varies from one word to a sentence. These events are identified by NLP systems. Information Retrieval (IR) systems, however, deal with text on a different level. For IR system, an event description corresponds to a collection of documents.

Systems participating in the Topic Detection and Tracking (TDT) tasks treated an event description as a collection of documents (Allan, Carbonell, Doddington, Yamron & Yang 1998). These are IR rather then NLP systems. According to the LDC annotation manual the TDT task:[3]

> A TDT ***event*** is defined as a particular thing that happens at a specific time and place, along with all necessary preconditions and unavoidable consequences. A TDT event might be a particular plane crash, or a single meeting, or a particular court hearing. An ***activity*** is a connected set of events that have a common focus or purpose, happening at a specific place and time; for instance, a campaign, or an investigation, or a disaster relief effort. For the purposes of TDT, a ***topic*** is defined as **an event or activity, along with all directly related events and activities**.

The problem of event definition in TDT was also complicated by the necessity of distinguishing between an event and a topic (Makkonen 2003). However, this distinction is made principally on the basis of specificity and targeted information retrieval rather than linguistic properties of the retrieved units. As Yang *et al* (1999) note, "[the] USAir-427 crash is an

[3] The complete annotation manual can be found here: `http://www.ldc.upenn.edu/Projects/TDT5/Annotation/TDT2004V1.2.doc`.

event but not a topic, and 'airplane accidents' is a topic but not an event." Initially, TDT systems did not aim at extracting information at less than the document level or structuring that information with semantic role annotation. Some current directions in TDT, such as new information detection (Allan, Gupta & Khandelwal 2001), operate on text passages smaller than the entire document. Such TDT systems are close in their interpretation of the term event to the systems that map events onto text snippets (Section 2.3).

2.6 Conclusions

In this chapter we show that, however intuitive the notion of 'event' might seem, different NLP applications use different definitions of what an event is. Not only is the exact meaning of an event in dispute, but also the extent of text used to realize an event (word, text snippet, relation with its semantic type, collection of documents). We showed the diversity of the event types currently identified in text. The identified events differ drastically in structure, amount of information they cover and even in the philosophy behind their definitions (i.e., events as grammatical objects and events as events in the world). We believe, that these various methods of event identification cannot and should not compete against each other. Instead, the combination of different event identification systems can lead towards creation of a very powerful end-to-end system which, given a query, would retrieve appropriate documents, identify pieces of text containing important information, assign time-stamps to these text pieces and finally, output a chronologically correct text (e.g., summary or an answer to a question). A pipeline for such a system is presented in Figure 2.1. The contributions of this dissertation come across different steps of this pipeline and connect events identified on different levels: from a collection of documents (TDT events) to verbs (verb class events). In Figure 2.1 we outline the portion that was implemented as part of the Columbia QA system. We present the detailed description of the Columbia QA system for answering event-related open-ended questions in Chapter 5.

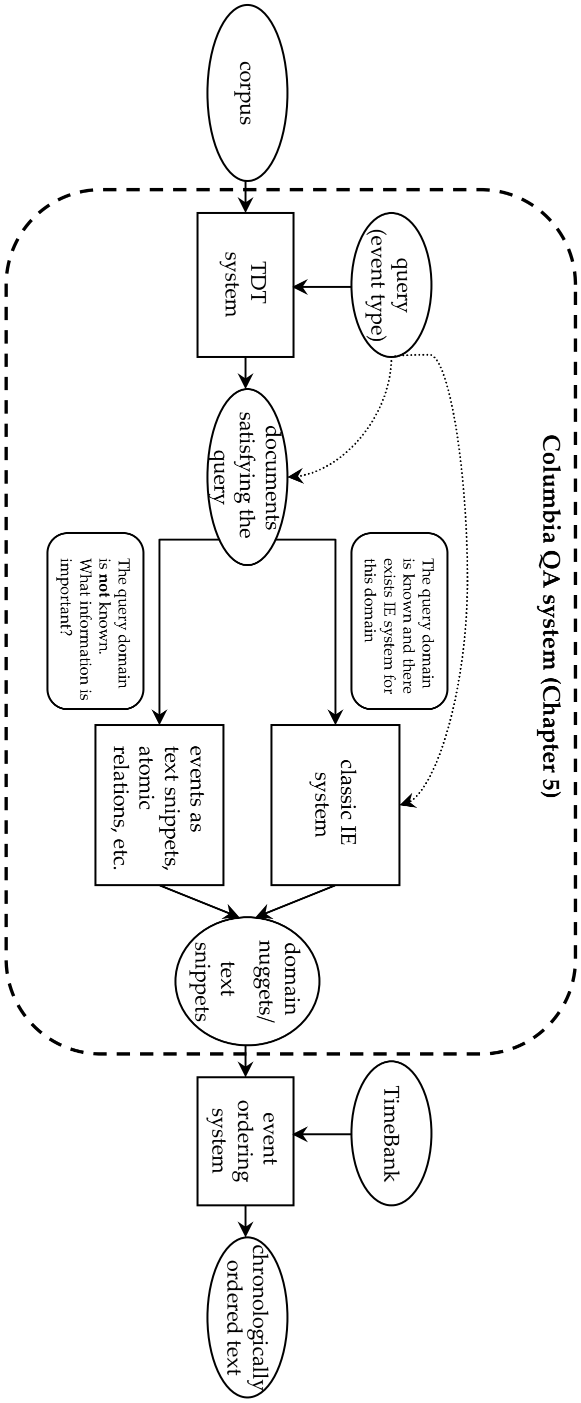

corpus
TDT system
query (event type)
documents satisfying the query
The query domain is known and there exists IE system for this domain
The query domain is **not** known. What information is important?
classic IE system
events as text snippets, atomic relations, etc.
domain nuggets/ text snippets
Columbia QA system (Chapter 5)
event ordering system
TimeBank
chronologically ordered text

Figure 2.1: Events pipeline. The portion of this pipeline that was implemented within the Columbia QA system for answering event-related open-ended questions is described in detail in Chapter 5.

Chapter 3

Shallow Semantic Network and Information Selection Formal Model

> Every person, all the events of
> your life are drawn there
> because you have them there.
> What you choose to do with
> them is up to you.
>
> Richard Bach

Many NLP applications deal with the task of information selection, where text snippets are extracted from an input document or collection of documents. For example, for the tasks of open-ended QA or summarization, multiple text snippets are selected and organized into a coherent text. For the general summarization task, this text should contain information that is judged as important for the input document collection; for the task of focused summarization or open-ended QA, the text snippets selected for the final output should target the information of interest that is either specified in the question or summary focus. Two major issues that should be resolved for information selection are:

- What text features should be used to identify the parts of input document collection that contain information of interest and thus, should be included in the final output?

- Once these parts of text are identified, in what order should they be added to the final output?

When the answer or summary consists of multiple separately extracted (or constructed) phrases, sentences, or paragraphs, additional factors influence the selection process. Obviously, each of the selected text snippets should individually be important. However, when many of the competing passages are included in the final output, the issue of information overlap between the parts of the output comes up, and a mechanism for addressing redundancy is needed. Current approaches in both summarization and long answer generation are primarily oriented towards making good decisions for each potential part of the output, rather than examining whether these parts overlap. Most current methods adopt a statistical framework, without full semantic analysis of the selected content passages; this makes the comparison of content across multiple selected text passages hard; usually, it is approximated by measuring the textual similarity between those passages.

Thus, most current summarization or long-answer question-answering systems employ two levels of analysis: a content level, where every textual unit is scored according to the concepts or information features it covers, and a textual level, where, before being added to the final output, the textual units deemed to be important are compared to each other and only those that are not too similar to other candidates are included in the final answer or summary. This second stage is often done using clustering. Clustering can be performed purely on the basis of text similarity, or on the basis of shared features that may be the same as the features used to select the candidate text units in the first place.

In this chapter we propose solutions for both aspects of the information selection process:

- We introduce a novel information feature (*atomic relation*) that can be used to capture information about the major concepts that are used in the event description. Atomic relations are constructed to connection these concepts among each other. Multiple atomic relations extracted from the input text can be analyzed within a shallow semantic network. This network successfully captures major facts described in the input text and can be used as a set of features for the process of information selection.

- We show how the procedure of passage selection can be mapped onto a well-defined complexity problem of set cover; thus, the best algorithm for the set cover problem is a prime candidate for the information selection process for such tasks as summarization and question-answering.

The rest of this chapter is organized as follows. In Section 3.1 we present an overview of related work on both aspects. We present a variety of information features that are currently used in information selection applications. We also describe methodologies that are used for scoring text units and ranking them in the order they should be output. As we are primarily interested in capturing information about the events described in the input text, we performed an annotation experiment where people marked text spans containing event descriptions. This experiment is described in Section 3.2. In Section 3.3 we describe the procedure that we developed for marking up novel features (atomic relations). In Section 3.4 we show how atomic relations can be later used to capture information about the events described in the input text. In Section 3.6 we discuss how the task of atomic relation identification corresponds to the task of Information Extraction. In Section 3.5 we report on the evaluation of the quality of the extracted atomic relations. In Section 3.7 we show how the problem of information selection can be formulated in terms of complexity theory. And, finally, in Section 3.8 we evaluate the quality of an information selection (summarization) system that uses atomic relations as information features and outputs text units following the greedy algorithm that corresponds to the set cover problem.

3.1 Related Work

The information selection process is driven by information features identified in the input text. These features are used to identify the text snippets that satisfy the selection criteria. For example, for a general summarization task, information features are used to identify text units (usually sentences) containing the most important information about the input document for single-document summarization and collection of documents for multi-document summarization. Many different kinds of information features have been for information selection tasks (e.g., summarization and QA).

There exist approaches that primarily use lexical features. For example, the most common content words are used as indicators of the main themes in the document. Sentences containing these words are scored using functions of their frequency counts (Edmundson 1968). Position of the words in text can also be an important factor. Words used in titles or document headings have higher importance then other words in the document and thus, good summarizers often include either titles or/and sentences containing words used in the titles and headings of the documents (Baxedale 1958, Edmundson 1968). Lead-based summarization baseline suggested in (Brandow, Mitze, & Rau 1995) is a widely used and very challenging baseline in the text summarization community. Lead-based summaries are constructed out of the first sentences of the input document. In two variations of lead-based summarization, summaries are constructed out of the first sentences from the first paragraphs of the document, or out of the first sentences from different documents for multi-document summarization. The presence or absence of certain cue-phrases like *significant*, *important*, *in conclusion* can be a good indicator of whether a sentence should or should not be included in the summary (Edmundson 1968). There are systems that analyze co-occurrence of concepts (Barzilay & Elhadad 1997, Hovy & Lin 1998).

Information selection features do not have to be lexical and directly observable within text snippets; they can represent abstract properties that particular text units may or may not satisfy, for example, status as a first sentence in a paragraph or, more, position in the source text (Baxedale 1958). Some summarization systems assume that the importance of a sentence is derivable from a rhetorical representation of the source text (Jones 1993, Ono, Sumlta & Miike 1994, Marcu 1997).

There is a separate class of summarization systems that leverage the fact that a summary should be produced for a collection of documents (McKeown & Radev 1995, McKeown, Barzilay, Chen, Elson, Evans, Klavans, Nenkova, Schiffman & Sigelman 2003). Such systems heavily rely on redundancy of various features present in the input text. This feature repetition can be also taken into consideration at the summary generation step (Barzilay & McKeown 2005).

Some summarization systems rely on relations sets that are manually identified as important for a particular domain and thus, should be included in the summary. Given a document collection corresponding to this domain, such systems use this set of relations to guide the summarization process (White, Korelsky, Cardie, Ng, Pierce & Wagstaff 2001, Radev & McKeown 1998).[1]

A document (or a set of documents) can be represented using a set of terms – known as a topic signature (TS) – that are highly correlated to the topic itself. Under this approach, each topic signature term is assigned an association weight that measures the relatedness of the term to the topic and thus, to the information rendered by the input document or collection of documents (Lin & Hovy 1997, Hovy & Lin 1998, Lin & Hovy 2000). The GISTexter summarization system (Harabagiu & Maiorano 2002) created templates for each input document collection. It used statistics over an arbitrary document collection together with semantic relations from WordNet. The created templates heavily depend on the topical relations encoded in WordNet. The template models an input collection of documents. Topic Themes (Harabagiu & Lăcătuşu 2005) used for multi-document summarization merge various arguments corresponding to the same semantic roles for the semantically identical verb phrases (e.g., *arrests* and *placed under arrest*).

After the features are identified in the input text, the process of information selection starts. The number of selected text snippets depends on how much information should be covered by the final output. The length of the output can also be bound by a predefined length (in words or characters). Thus, it is important to be able to rank sentences and output them in such an order so that most important text snippets are selected ahead of less important text-snippets. Summarization can be treated as a classification problem (Kupiec, Pedersen & Chen 1995, Teufel & Moens 1997), where each sentence is assigned a probability; the sentence with the highest probability is output first, probabilities for the remaining sentences are recalculated and the process is repeated until the output text reaches the target length. Goldstein *et al* (Goldstein, Mittal, Carbonell & Callan 2000) output sentences to minimize information redundancy in the output. The information selection process in the DefScriber QA system (Blair-Goldensohn, McKeown & Schlaikjer 2004) is guided by a pre-defined definition structure. Zhou *at al* (Zhou, Ticrea & Hovy 2004) output only those

[1] Appendix A has an example of "terrorist attack" template used for summaries generation in (Radev & McKeown 1998).

sentences that are classified as biographical according to seven pre-defined features. Information retrieval techniques for automatic generation of semantic hypertext links focuses were used for automatic text summarization in (Salton, Singhal, Mitra & Buckley 1997).

Thus, to create a successful information selection application two issues should be addressed:

- input text should be broken into text units each of which should be marked according to the presence/absence of the identified information features;
- text units should be ranked according to the order they will be added to the final output.

This two-stage process corresponds to the two-stage human production of abstracts: the *analytical* stage in which the salient facts of the text are obtained and condensed and the *synthetic* stage in which the text of the abstract is produced (Molina 1995)

3.2 A Study of Event Annotation

As noted in Chapter 2, there is huge diversity in both the structure and length of what is identified as event descriptions in text, from one verb for linguists to a collection of documents for the TDT task. We show that different event identification standards should not compete; rather, as they identify events on different levels of text they should leverage each others' output.

The choice of the event description approach depends on the initial task. For example, for factoid QA systems the type of the input question corresponds to a particular questions type (e.g., the *When was X born?* question can be treated as a question of the Date-Of-Birth type). Thus, factoid questions can be often mapped to relation types that are similar to the event description relations studied by classic IE systems (Sections 2.4 and 3.6.2). The IE approach for event descriptions can be also used for the summarization systems that are designed to create summaries within a particular domain (White et al. 2001, Radev & McKeown 1998). At the same time, there exist summarization systems that are designed to summarize the latest news about what is going on the world. Such summarization systems typically treat event descriptions as text snippets (Section 2.3). Thus, before going any further we describe our task. The goal of this task will guide the choice of event description mechanism that we are going to use.

In the current work we are interested in selecting information about events described in the input text (collection of documents). Thus, obviously, the IR approach for event description does not suit our task as the granularity of our descriptions should be smaller than a document collection. TimeML specification is more suited for labeling event-related words and phrases mentioned in text. IE event descriptions encode relations among various text entities that are important for the input text. Text snippets do not encode relations, however, they can be used in the cases where there is no event description type that can be identified in text (e.g., the input text is about a new domain for which no important relations have been defined yet). Thus, choosing among the TimeML specification, text snippets, or IE approach for describing events is not an easy task. To identify which of the event description approaches suites our task best we run an experiment where we ask human annotators to mark up events mentioned in text.

Given a document collection as input our system should identify what length of text snippets capturing information about events is more appropriate for our task: a single word, a simple clause, or a sentence. Thus, before going any further we describe an experiment, the major goal of which is to obtain a definition of events that can be used in a system for the automatic detection and extraction of events from a document.

We first conducted a study of text annotation for event information by asking a number of computer science graduate students (mostly in computational linguistics) to mark text passages that describe events in news stories. We deliberately provided no definition of *event* for this study, to see if the respondents would naturally converge to an operational definition (as evidenced by high agreement on what they marked). The annotators were given 13 news articles randomly selected from the DUC-2001 (Document Understanding Conference) corpus. The texts varied in length from 15 to 60 sentences. Five of the thirteen texts were annotated by two participants in the study. In addition to checking for agreement between the annotators and anecdotal evidence of the difficulty or ease with which they could label events, our study had two further aims: To determine what text ranges, in the

One (full) sentence	*Several sentences*	*Less than a sentence*	*Simple clause extracted from a longer sentence*
95 (50%)	27 (14%)	22 (11%)	46 (24%)

Table 3.1: Length of text snippets covering event descriptions.

absence of instructions on the length of what they should mark, people tend to favor as the appropriate descriptors of a single event; and to gather evidence of features that occur with high frequency in the marked passages and could be automatically extracted by an automated system simulating the human annotators.

3.2.1 Length of Marked Text Passages

While the annotators disagreed on what text pieces to select as event descriptions, they exhibited more agreement on how long these pieces should be.

Table 3.1 shows the distribution of the lengths of the pieces that were marked, among clauses, sentences, and longer units. We also consider a variation of approximately one-sentence long units: a sentence minus one short prepositional phrase (denoted `Less than a sentence` in Table 3.1). This column represents cases where the annotator thought that the omitted phrase was not crucial to the description of the event. To our surprise, we found that in most of these cases the phrase omitted was temporal, as in

> *On Tuesday, the United States and Soviet Union conferred in Washington on putting an anti-Iraq naval blockade under a United Nations umbrella.*

where the “on Tuesday” part was not included in the event description. This is contrary to most event structure theories that assume a central role for time in events. We hypothesize that this may be due to the fact that the events described in the articles occurred in 1988–1991, and thus an exact order of the day in week ten years ago may not have been considered important by the annotators.

According to Table 3.1, the simple clause is really the minimal unit representing marked-up events (noun phrases such as *war* or *earthquake* were never marked as events). However, twice as many full sentences as simple clauses were marked as events, and an additional 11% of the marked regions were almost full sentences. What is remarkable from Table 3.1 that in half of the cases annotators chose a sentence-length unit as a text region for an event. We therefore conclude for our task, a full sentence appears to provide the most reasonable scope for an event description.

In (Filatova & Hovy 2001) the syntactic counter part for the intuitive understanding of an event is equal to a simple clause containing a tensed verb. Such simple clauses are called event-clauses. According to our evaluation, the simple clause is really the minimal unit covering an event description (noun phrases such as *war* or *earthquake* were never marked as events). But the percentage of the simple clauses marked as events is two times less than the percentage of the sentences marked as events. In addition, we assume that those events that are represented by several sentences can be treated as several events united together and we can also stretch to the sentence-level those events that are represented as by less than a sentence. Taking into consideration these observations plus all the problems that were described in (Filatova & Hovy 2001) for breaking sentences into simple clauses, we decided to take a sentence as a region covering an event description.

3.2.2 Text Features in Marked Passages

We analyzed the passages marked as event descriptions looking for text features that could be included in an automated event detection system. Naturally, the verb itself often provides important information (via tense, aspect, and lexical properties) about the event status of a clause or sentence. In addition, the following features are correlated with the presence of events: **Proper nouns** occur more often within event regions, possibly because they denote the participants in events. In contrast, **pronouns** are less likely to occur in event regions than in non-events. As expected, the presence of **time phrases** increases the likelihood of a text region being marked as an event description.[2] **Cardinal numbers** were another lexical class strongly associated with events. This can be attributed to the fact that numbers are often given when new important information is presented; they condense information and typically accompany factual rather than subjective sentences, which are more likely to be associated with event descriptions.

[2] Unless, the time phrase is a name of a day of week (e.g., Monday, Tuesday, Wednesday, etc.) for an event that happened several years ago.

Absence of Pronouns and Presence of Proper Names Events have participants, and it is likely that important events will mention the participants by name. Often a sentence introduces an event with named participants, with subsequent related sentences elaborating on particular aspects of the event; pronouns are more likely to be used in secondary sentences or clauses. Text is typically organized by topic and time, and the use of proper names often signifies a switch to a different topic or, when time changes, to a different time, as observed by (McCoy & Strube 1999).

Cardinal Numbers Most of the simple clauses containing figures (e.g. Richter scale magnitude of an earthquake, amount of money spent on something, number of people killed during some accident, etc.) were labeled as important events. This can be attributed to the fact that numbers are often given when new important information is presented; they condense information and typically accompany factual rather than subjective sentences, which are more likely to be associated with event descriptions. For example,

> *First group of 230 evacuees was assembling in Baghdad today.*
>
> *Some 13,000 people were injured and more than a half-million were left homeless in the Armenian earthquake.*

65% of the sentences containing cardinals were chosen as events and such sentences tend to be important events. Moreover, the amount of cardinals per sentence marked as event is more than the amount of cardinals per sentences not marked as event.

Time Phrases Most event theories associate events with definite points or periods of time. Therefore, it is likely to expect that the presence of a temporal phrase (a prepositional phrase, noun phrase, or simple clause indicating time) increases the likelihood of a text region being marked as an event description.

An interesting exception to this positive association between event descriptions and time phrases is that occasionally, as we mentioned above, annotators failed to include a time phrase in the event region although they included the rest of the sentence. This always happened with time phrases that are not very specific, as in the case when the article is read with a significant delay after the event took place (e.g., "on Tuesday"). In contrast, more specific time phrases such as a year or a month/year combination were always included as part of the event description.

> *USS Vincennes shot down the Airbus on July 3, 1988*

3.3 Identifying Features for Event Description

Drawing from our event annotation study described in Section 3.2, we decided on an algorithm for detecting, extracting, and labeling atomic relations that is based on the features that were more strongly correlated with event regions. In this section we show that atomic relations can be successfully used as information selection features, especially when input text (document or a collection of documents) describes an event.

According to the above analysis, event regions are contained within a sentence. Thus, we anchor relations on their major constituent parts (named entities for people and organizations, locations, and time information)[3] and require at least two such major elements in a sentence before considering extracting a relation. The procedure for extracting atomic relations is the following:

- We analyze a collection of documents clustered on a specific topic.
- We take the sentence as the scope of a relation. Our algorithm ignores sentences that contain one named entity or none.
- We extract all the possible pairs of named entities.
- For each pair of named entities we extract all the words that occur in between the elements of the relation. These are extracted together with their part of speech tags which we get with the help of Collins' (1996) parser.[4]
- Out of all the words that occur in between elements of named entity pairs, we are now interested only in those which are either non-auxiliary verbs or nouns which

[3] All these major elements can be retrieved with a named entity tagger; for the experiments described in this Chapter we use BBN's IdentiFinder (Bikel, Schwartz & Weischedel 1999).

[4] As in the case with named entity recognition, any other POS tagger or syntactic parser can be used to mark-up verbs and nouns in the input text.

are hyponyms of *event* or *activity* in WordNet (Miller 1995).We call these words *connectors*.

- For each relation we calculate how many times it occurs, irrespective of the connectors.
- For each connector we calculate how many times this connector is used in a particular relation.

Triplets identified by the above algorithm are called *atomic relations*.

Our hypothesis is that if named entities are often mentioned together, these named entities are strongly related to each other within the topic from which the relation was extracted. Although our method can be applied to a single text (which by itself assures some topical coherence), we have found it beneficial to extract atomic relations corresponding to the described event from a set of related articles. Such sets can be created by clustering texts according to topical similarity, or as the output of an information retrieval search on a given topic. Following the above procedure we create a list of relations together with their connectors for a set of documents.

Out of all the relations we keep only the most frequent ones and we also eliminate those relations that are not supported by high frequency connectors (both of these parameters are adjustable and are determined empirically).

The described algorithm is a symbiosis of the TimeML specification, text snippets, or IE approach for describing events. We are interested in learning important relations among text entities. At the same time, we do not use any predefined IE-like event description types. Rather, we use non-auxiliary verbs or nouns which are hyponyms of *event* or *activity* as the connectors linking text entities. We also collect only those relations that appear within text snippet of a particular length (i.e., sentence).

3.4 Shallow Semantic Network

We created a system that automatically identifies atomic relations. We tested our system on a subset of the topics provided by the Topic Detection and Tracking Phase 2 research effort (Fiscus, Doddington, Garofolo & Martin 1999). The topics consist of articles or transcripts from newswire, television, and radio (the New York Times, Associated Press,

THING: China Airlines Flight 676 from Bali to Taipei crashes
PLACE: Taipei, Taiwan
WHEN: February 16, 1998
TOPIC EXPLICATION: The flight was from Bali to Taipei. It crashed several yards short of the runway and all 196 on board were believed dead. China Airlines had an already sketchy safety record. This crash also killed many people who lived in the residential neighborhood where the plane hit the ground. Stories on topic include any investigation into the accident, stories about the victims/their families/the survivors. Also on topic are stories about the ramifications for the airline.

Figure 3.1: Official description of *China Airlines crash* topic.

CNN Headline News, ABC World News Tonight, PRI The World, and Voice of America English News Service). In this corpus, the amount of text describing a topic varies from 1 up to 3872. We used 70 of the 100 topics, those containing more than 5 but less than 500 texts. Since human annotators created these topical clusters in a NIST-sponsored effort, we can be assured of a certain level of coherence in each topic. In this manner, we can concentrate on the benefits or shortcomings of our algorithm rather than on issues related to the retrieval of on-topic texts.

TDT provides descriptions of each topic that annotators use to select appropriate documents by issuing and modifying IR queries. The official description of one topic ("China Airlines crash") is given in Figure 3.1. Table 3.2 shows the top 10 pairs of named entities extracted from the topic at the first stage of our algorithm (before considering connectors). The normalized relation frequency is calculated by dividing the score of the current relation (how many times we see the relation within a sentence in the topic) by the overall frequency of all relations within this topic.

It is clear from the table that the top relations mention the airline company whose plane crashed (*China Airlines*), where the crash happened (*Taiwan*, *Taipei*, *International Airport*), where the plane was flying from (*Bali*), and when the crash happened (*Monday*). Interestingly we obtain a clique for the three elements *China Airlines*, *Taiwan*, and *Taipei*. Let us analyze the connectors for the three pairs among these three elements (Table 3.3). The normalized connector frequency is calculated by dividing the frequency of the current connector (how many times we see this connector for the current relation) by the overall frequency of all connectors for the current relation.

China Airlines is linked to both *Taipei* and *Taiwan*, and the lists of connectors are similar

Relation Frequency	First Element	Second Element
0.0212	China Airlines	Taiwan
0.0191	China Airlines	Taipei
0.0170	China Airlines	Monday
0.0170	Taiwan	Monday
0.0170	Bali	Taipei
0.0148	Taipei	Taiwan
0.0148	Bali	Taiwan
0.0148	Taipei	Monday
0.0127	Bali	Monday
0.0127	International Airport	Taiwan

Table 3.2: Top 10 named entity pairs for the *China Airlines crash* topic.

enough for our system to merge the two extracted events to one. On the other hand, there is no event connector linking *Taipei* and *Taiwan*. Our system assumes that this relationship is a static one (indeed, Taipei is the capital of Taiwan), and drops this candidate event. The final output is shown in Table 3.4. The connectors output by the system highlight the major event linking *China Airlines* and {*Taiwan*, *Taipei*}, that is, the crash. The importance of these connectors is also verified by calculating the relative connector frequencies for the entire topic, irrespective of the specific entities involved (Table 3.5).

From the above table it is clear that *to crash* is the major event not just for the top relations but for the whole topic. And it proves that the relations that are on top of the relations list are really the most important relations for the topic.

Finally, we factor in topic specificity for the extracted events and sort the extracted events . Tables 3.6 shows the most and least specific named entity pairs for this topic. The less specific entries correspond to generic relationships (e.g., there are only seven week days), relationships totally independent of the topic (e.g., *Bill Hazard* reports from *Washington*), and some are related but not limited to this topic (e.g., *China* and *Taiwan* have a long relationship separate from this crash, resulting in their mention in other topics as well). In our example, the top event of Table 3.4 is specific to this topic, but other events further

Relation	Connector	Connector Frequency
China Airlines – Taiwan	crashed/VBD	0.0312
	trying/VBG	0.0312
	burst/VBP	0.0267
	land/VB	0.0267
China Airlines – Taipei	burst/VBP	0.0331
	crashed/VBD	0.0331
	crashed/VBN	0.0198
Taipei – Taiwan	–	–

Table 3.3: Top connectors for three of the relations in Table 3.2.

First named entity	Second named entity	Connectors
China Airlines	Taiwan; Taipei	crashed/VBD
		trying/VBG
		burst/VBP
		land/VB
		killing/VBG

Table 3.4: Final event output for the relations of Table 3.2.

down in the list (such as the *China–Taiwan* one) are deemed non-specific and pushed further down or removed from the output.

Thus, our definition of an event is algorithmic rather than semantic. For us an event is a triple of elements where two named entities are linked between each other through a verb or a noun which is a hyponym of *event* or *activity* in WordNet (Miller 1995). These triples are constructed based on simple co-occurrences of the three elements in the input sentences and do not hold any syntactic or semantic dependencies.

We close this section with a comment on the anchor points used by our algorithm. Such anchor points (by default, named entities) are necessary in order to limit the amount of relations considered. We chose named entities on the basis of our analysis of events marked by people (Section 3.2). However, the system is adaptable and the user can specify

Connector	Frequency across topic
crashed/VBD	0.0189
burst/VBP	0.0107
trying/VBG	0.0092
land/VB	0.0079

Table 3.5: Top connectors across the entire *China Airlines crash* topic.

Relation	Specificity
China Airlines – Monday	1.0000
Taiwan – Monday	1.0000
Bali – Taipei	1.0000
Beijing – Tuesday	0.5681
Bill Hazard – Washington	0.4815
Wednesday – Monday	0.2448
Tuesday – Monday	0.1922
China – Taiwan	0.1582
CNN – New York	0.0850

Table 3.6: Pairs which are and are not specific for the *China Airlines crash* topic.

additional words or phrases that should be used as anchor points. In this example, it makes sense to extract information involving the passengers of the plane. If the word *passengers* is submitted to the system, then the third from the top events extracted will refer to the deaths of the passengers, as shown in Table 3.7.[5] In Figure 3.2 we show a graphical representation of atomic events where multiple Named Entities linked through common event labels. This figure shows how atomic relations can be merged into a shallow semantic network.

[5] 197 is the correct number and it was used more often than the other two numbers which were given in the early articles describing this crash when the exact numbers were not clear yet.

Event elements	Verbs	Nouns
Taiwan – passengers	killing/VBG carried/VDB	197/CD 196/CD 182/CD

Table 3.7: Event extracted for the noun "passengers" from the *China Airlines crash* topic.

3.5 System Evaluation

3.5.1 Methodology

To evaluate our system, we randomly chose 9 topics out of the 70 selected TDT-2 topics. For each of these topics we randomly chose 10 texts, and ran our system on these 10 texts only, producing a ranked list of events with verb and noun labels, as described in Section 3.3. We then gave the texts and the top 10 events in the system output for a given topic to a volunteer evaluator (a graduate student in computational linguistics). Each evaluator processed exactly one topic.

Our subjects were asked to first read the texts[6] and then provide a numerical score for the system in the following areas:

1. Whether the named entities in the events extracted by our system are really related to each other in the texts. A separate score between 0 and 1 was given for each extracted event.

2. Whether the extracted relations between named entities, if valid, are also important. Again a 0 to 1 score was assigned to each extracted event.

3. Whether the labels provided for a (valid) event adequately describe the relationship between the named entities.

For these three questions, the evaluators gave a separate score for each extracted event. Rather than asking our annotators to mark the extracted relations as either relevant or irrelevant, we asked them to assign to each relation a degree or relevance on a scale between

[6] Which was the reason we limited the number of texts per topic to 10.

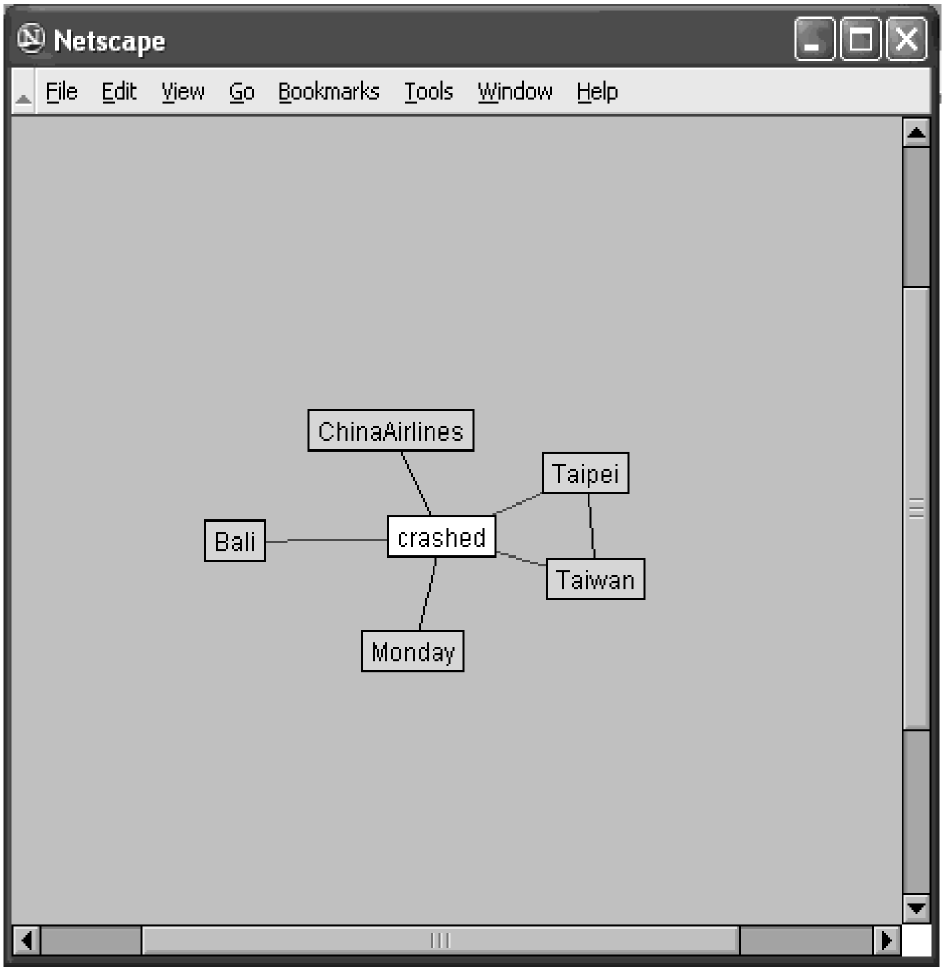


Figure 3.2: Screen shot of the system output.

0 (irrelevant) and 1 (relevant). The choice of the optimal answer format is an interesting question that is being extensively studied in sociology, political science, marketing research, etc. (Hensler & Stipak 1979, Dolnicar & Grün 2007). We chose to use a scale rather than a set of values as the extracted relations did not have any semantic tags assigned to them and thus, were rather difficult to evaluate as completely relevant without additional assumptions.

In addition, we asked evaluators to enumerate important events that the system missed, and provide a subjective rating between 0 and 1 on how closely related the articles in their set were.

Question	Average rating	Percentage non-zero	Percentage above 0.5
Link quality	0.7506	92.22%	74.44%
Importance	0.6793	95.00%	62.87%
Label quality	0.6178	90.91%	51.09%

Table 3.8: Evaluation scores for our system. Importance and label quality measured only on extracted relations of reasonable quality (with link quality score above 0.5, 75% of the total extracted events).

3.5.2 Results

Table 3.8 shows the scores obtained during the evaluation. We report the average rating our system obtained on each of the three questions, across both the ten extracted events in each set and the nine evaluators/topics. We also report the percentage of extracted events that received a non-zero score and a score above 0.5.

We note that the easiest task for the system is to find valid relationships between named entities, where we obtain about 75% precision by either the average score or the number of scores above 0.5. Next comes the task of selecting important links, with precision of 63–68%. The hardest task is to provide meaningful labels for the events; we succeed in this task in slightly more than half of the valid extracted events, or approximately 40% of the total extracted events.

Since the difficulty of the task is correlated with the coherence of the document sets being analyzed, we observed significant differences in the scores between topics. In some cases our system obtained scores above 70% or 80% in all three questions. In two cases, the scores were below 20%; in one of those, the documents covered a very wide range of events (many different events related to the Israeli-Palestinian peace negotiations), while the other topic dealt with an earthquake in Afghanistan. In the latter case, our system, which looked by default for named entities, could not extract enough event relations as the participants of the this event were expressed not by named entities but, rather, by general nouns and noun phrases. Regardless, our system overall extracted at least somewhat useful information, as manifested by the fact that 90% of the reported events received non-zero scores.

3.6 Atomic Relations and Information Extraction

The system described in Section 3.3 identifies the relations that can be used for identifying what are the most important events described in the input collection of documents. Systems dealing with the task of Information Extraction (IE) (Grishman 1997) have similar goals. In this section we give an overview of the existing IE system and describe similarities and differences between IE systems and atomic relations.

3.6.1 History of Information Extraction (IE)

The ability to automatically select elements of interest from a free text was realized as important very early on. In 1987 the first Message Understanding Conference (MUC) was held. As input, MUC systems got types of information defined as important for a particular domain. As output, MUC systems produced text snippets satisfying the information types of interest. The systems developed for MUC-1987 and MUC-1989 focused on documents from naval communication domain; the systems developed for MUC-1991 and MUC-1992 focused on documents from terrorism domain, specifically on documents describing terrorist attacks in Latin America; the systems developed for MUC-1993 focused on documents about joint ventures and microelectronics; the systems developed for MUC-1995 focused on documents about management changes; and the systems developed for MUC-1993 focused on reports about launches and air crashes.

Systems participating in MUC competitions developed a variety of pattern-generation techniques for selecting important text pieces for a particular domain (Muslea 1999). The description of the classical IE task within the MUC framework was finalized as:

> Information Extraction is the process of identifying relevant information where the criteria for relevance are predefined by the user in the form of a template that is to be filled. Typically, the template pertains to events or situations, and contains slots that denote who did what to whom, when, and where, and possibly why. The template builder has to predict what will be of interest to the user and define its slots and selection criteria accordingly. (Hovy, Ide, Frederking, Mariani & Zampolli 1999)

Thus, it became clear that the IE task, as it was developed within the MUC framework, was inherently a domain-specific task (Cardie 1997).

The performance of MUC systems varied greatly for different tasks. The best performance was achieved for the task of named entity tagging with the F-measure of 96.43% for the best performing system. The most difficult task was the scenario template (ST) task where the systems had to fill-in slots in a manually created domain templates. These slots corresponded to the information about entities, relations, etc. The F-measure for this task was in the range of 35.60% – 41.18%.[7]

A recent initiative on information extraction systems evaluation, the ACE evaluation, is a successor of MUC evaluations. As in the MUC evaluations, systems participating in the ACE evaluation are given sets of pre-defined named entity, relation/event types that should be extracted.[8] ACE includes both Entity Detection and Tracking (EDT) and Relation Detection and Characterization (RDC). EDT is broadly comparable with the MUC Named Entity (NE) task, while RDC is broadly comparable with the MUC template elements task. However, both ACE tasks are more challenging than their MUC forerunners (Maynard, Bontcheva & Cunningham 2003).

Rapid development of the Internet in the past decade had a two-fold influence on the field of NLP. On the one hand, NLP systems got access to a very large corpus (probably the largest existing corpus); however, it has been noted that this corpus is very noisy and the data retrieved from it requires filtering. Moreover, the necessity of dealing with such a huge corpus demonstrated the importance of scalability for NLP applications. It became clear that processing terabytes of data is time-consuming and IE patterns were shown to be very helpful for quick processing of large volumes of data (Shinyama & Sekine 2006, Banko, Cafarella, Soderland, Broadhead & Etzioni 2007).

3.6.2 Overview of IE Systems

All the current IE systems can be divided into two large groups:

[7] The presented F-measure values were reported in (Marsh & Perzanowski 1997).

[8] The description of the ACE task as of July 1, 2005 can be downloaded from `http://projects.ldc.upenn.edu/ace/docs/English-Events-Guidelines_v5.4.3.pdf`.

1. systems that require as input a pre-defined set of named entity types, or relation/event types that should be extracted;
2. systems that do not require any preliminary information on the data type to be extracted, rather, their goal is to extract all possible meaningful relations present in the input corpus.

The first type of systems borrow a lot of techniques from the classic MUC-style IE systems. Such systems are described in Section 3.6.2.1. The systems that belong to the second type are described in Section 3.6.2.2.

3.6.2.1 Extraction Systems Dealing with Pre-Specified Relation Types (Classic IE Systems)

In this section we describe IE systems that deal with MUC or similar tasks of extracting entities, relations, or events according to the pre-defined specifications.

First IE system relied on manually created dictionaries and extraction pattern sets to identify and extract relevant information. The manual process of creating dictionaries and extraction pattern sets was extremely time-consuming. Moreover, the created valuable data was not portable across different domains. One of the first attempts to address this knowledge-engineering bottleneck with the AutoSlog system (Riloff 1993). The AutoSlog system automatically created extraction patterns that could be used to construct dictionaries of important elements for a particular domain. AutoSlog required a text where the elements of interest were manually tagged only for the training stage. Another input requirement for AutoSlog was a set of syntactic rules that could capture the tagged elements. However, it must be noted that though it was shown the initial set of syntactic rules that worked well for one domain did not necessarily work well for a different domain, these syntactic patterns were very generic.

AutoSlog-TS (Riloff 1996) attempted to overcome the necessity of having a hand-labeled input. As input it required only preclassified texts and a set of generic syntactic patterns. The procedure devised for AutoSlog-TS captured a lot of relevant syntactic patterns that could be used within an IE system, the precision of some of those patterns, however, was low. Thus, before using the learned patterns within an IE system, manual filtering or irrelevant patterns was performed.

The PALKA system (Kim & Moldovan 1995) learned extraction patterns that were expressed as *frame-phrasal pattern structures* (FP-structures). An example of an FP-structure is presented in Figure 3.3. Each slot in the meaning frame defines an item-to-be-extracted together with the semantic constraints associated with this item. The phrasal pattern represents an ordered sequence of lexical entries and/or semantic categories taken from a predefined concept hierarchy. The FP-structure combines the meaning frame and the phrasal pattern by linking the slots of the former to the elements of the latter. PALKA FP-structures are similar to the Case Frames (Riloff & Schmelzenbach 1998) described in Section 4.1 (Chapter 4). As input, PALKA requires lists of key-words associated with the FP-structure domain and semantic lexicons generated for the FP-structure domain. PALK con-
straint

```
FP-structure = MeaningFrame + PhrasalPattern
Meaning Frame: (BOMBING agent:      ANIMATE
                        target:     PHYS-OBJ
                        instrument: PHYS-OBJ
                        effect:     STATE

Phrasal Pattern:  ((PHYS-OBJ) was bombed by (PREP))

FP-structure:
    (BOMBING   target:    PHYS-OBJ
               agent:     PERP
               pattern:   ((target) was bombed by (agent))
```

Figure 3.3: Example of an FP-structure.

The Rapier system (Califf & Mooney 1998) followed a supervised learning approach and used a set of pairs: a text and a filled in template corresponding to this text (the initial experiment was run for the job advertising domain). The patterns were created to capture the text snippets which could be good fillers for the input template.

Systems created specifically for MUC tasks were not the only IE systems dealing with retrieving relations of pre-defined types. For example, the DIPRE (Dual Iterative Pattern Expansion) system (Brin 1998) was proposed as an approach for extracting a structured relation (or table) from a collection of HTML documents. The method works best in an environment like the World-WideWeb, where the table tuples to be extracted will tend to appear in uniform contexts repeatedly in the collection documents (i.e., in the available HTML pages). DIPRE exploits this redundancy and inherent structure in the collection to extract the target relation with minimal training from a user. DIPRE's successor, the Snowball system (Agichtein & Gravano 2000), used bootstrapping technique to get a set of tuples corresponding to a particular relation and patterns capable of capturing these tuples. As input, Snowball required a small set of seed tuples that were representatives of the relation under analysis.

Techniques introduced by classic IE systems were successfully used for the task of factoid question answering. For example, the system implemented by (Ravichandran & Hovy 2002) is based on the same principle as Snowball: a hand-full of pairs corresponding to the input question type are submitted as an input for the system; an output of the system is a set of patterns that could be used to find answers to the questions of the same type. A QA system seeking long elaborate answers can also be considered an information extraction system in the sense that it extracts sentences that contain information of interest to the user; usage of extraction patterns for such systems also proved to be useful (Blair-Goldensohn, Evans, Hatzivassiloglou, McKeown, Nenkova, Passonneau, Schiffman, Schlaikjer, Siddharthan & Siegelman 2004).

Approaches similar to those used in classic MUC-style IE systems are also used for creating large data repositories that contain tables of relation examples corresponding to a particular relation type. Several such systems and associated data repositories containing binary relations have been created and evaluated in the last several years. Hearst (1992) examined the usage of lexical patterns to extract hyponym relations from text. Berland and Charniak (1999) extracted "part-of" relations. Snowball IE system (Agichtein & Gravano 2000), given a set of seed tuples corresponding to a particular relation uses two-step algorithm for getting tuples corresponding to the same relation. Fleischman *et el* (2003) created a table containing relations for answering "Who is X?" factoid questions. Barbara Rosario and Marti Hearst (2004) use generative models to identify the patterns to capture the relations annotated in the training corpus. Other researchers (McDonald, Pereira, Kulick, Winters, Jin & White 2005) concentrate on identifying those sets of noun phrases which correspond to the input n-ary relation. To do this, McDonald *et al* use the results of a named entity tagger which marks those named entities which have the same type and the input n-ary relation and then use graphical models to extract the appropriate relations.

Several systems use methodologies for extracting instances corresponding to a particular relation type from text. For example, Mann (2002) explores the idea of fine-grained proper noun ontology. This ontology can be used to enhance WordNet and is shown to be useful for a QA task. The KnowItAll IE system (Etzioni, Cafarella, Downey, Kok, Popescu, Shaked, Soderland, Weld & Yates 2004) acquires instances of cities, states, countries, actors, films, etc. KnowItAll uses bootstrapping and as an input it requires a set of generic rule patterns and several seed tuples corresponding to a specific relation type (for example, <*countries* such as *Great Britain*, *France*, and *Germany*>).

3.6.2.2 Extraction Systems Dealing with Extraction of Relations of All Types from the Input Corpus

The classic MUC-style IE systems described in Section 3.6.2.1 require as input a set of relations and entity types that should be extracted. Recently there has been a lot of interest in creating data repositories for all possible relations present in the input corpus (Hasegawa, Sekine & Grishman 2004, Shinyama & Sekine 2006, Sekine 2006, Banko et al. 2007). The main feature of such systems is that they do not require as input pre-defined types of data of interest. Rather, the goal of such systems is to extract from the input corpus all the useful relations corresponding to a variety of relation types. These relations, however, are usually difficult to categorize. Furthermore, many relations contain loosely connected entities, capturing nothing more than a mere frequent co-occurrence of these entities in the text.

As already noted, the systems described in Section 3.6.2.1 are not easily portable to get tuples corresponding to new relations or question types. Thus, many recent IE systems have targeted the task of getting all possible relations (tuples and patterns capturing these tuples) from large corpora (for example, from Web). This task is called *Open or On-Demand IE (OIE)* or *Unrestricted Relation Discovery* and comes as a continuation of our work on atomic relation extraction.

Hasegawa *et al* (2004) describe a procedure for discovering significant relations embedded in documents. The system analyzes pairs of named entities that co-occur within one sentence with the context around these named entities. The final evaluation is performed on two types of relations PERSON-GPE and COMPANY-COMPANY and was centered around the IS-A relation. For example, PERSON *is a president of* GPE. The On-Demand Information Extraction system (Sekine 2006) and the Unrestricted Relation Discovery introduced in (Shinyama & Sekine 2006) introduce a further development of this idea. The goal of these systems is: given a set of related documents (about an event or similar events) discover all possible relations from these documents and present them as tables. To discover these relations, the system introduced in (Shinyama & Sekine 2006) uses a logical feature structure called GLARF (Grammatical and Logical Argument Representation Framework) (Meyers, Grishman, Kosaka & Zhao 2001). GLARF has an ability to regularize several linguistic phenomena such as participial constructions and coordination, and an output structure can be easily converted into a directed graph that represents the relationships between each word, without losing significant information from the original sentence. Then, automatically obtained relations are clustered according to the similarity of their contexts. Each relation cluster is listed within one table where the number of table columns is equal to the number of entities in the relation and the titles of table columns are set to be equal to the type of the relation entities. The *consistency* of the tables that have at least three rows is evaluated.

> an evaluator looks at every row and its source article, and tries to come up with a sentence that explains the relations among its columns. The description should be as specific as possible. If at least half of the rows can fit the explanation the table is considered "consistent."

In the evaluation procedure of atomic relations described in Section 3.5 we also provided our annotators with the source documents so that they could judge whether an atomic relation captures a valid relation or not. However, we do not cluster atomic relations and thus, evaluate each relation separately, while Shinyama *et al* (Shinyama & Sekine 2006) perform a collective evaluation of a set of similar relations.

The On-Demand Information Extraction system (Sekine 2006) uses sub-trees of dependency trees which are relatively frequent in the retrieved documents compared to the entire corpus. The relations for the evaluation are extracted using 31 queries constructed by the authors. These queries are formed to satisfy the ACE event types.

The TextRunner system (Banko et al. 2007) goes beyond analyzing a collection of related documents. Its goal is to get all possible relations from the web. The extracted relations are then evaluated and judged as either salient or not. Many of the relations captured by TextRunner are of high quality. It is difficult, however, to infer the precise semantics of the extracted relations. Furthermore, a lot of computation time is spent on capturing relations that are either too general or difficult to judge. For example, the data analysis presented in (Banko et al. 2007) shows that out of more than 11 million tuples less then 10% of the tuples can be evaluated and used later in other tasks (including QA); the rest of the tuples are either too abstract or not well-formed.

The task of *Open or On-Demand IE (OIE)* or *Unrestricted Relation Discovery* is new and there does not exist a standard procedure for measuring the quality of the extracted relations. Moreover, it is not clear how to measure whether all valuable relations were extracted by any of the described systems. We believe that the quality of the relations can be judged through the tasks where these relations are used. In this dissertation we use atomic relation as information selection features for the tasks of multi-document summarization (Chapter 3) and open-ended QA (Chapter 5).

3.6.3 Comparison of Atomic Relations and IE Systems

Atomic relations described in this chapter connect named entities through an action. Thus, atomic relations capture ternary patterns of the form

$$\langle NamedEntity1 - Verb - NamedEntity2\rangle.$$

According to the described procedure, the system extracting atomic relations does not require as input a set of pre-defined relation types to be extracted. At the same time, the system does not assign exact semantic types to the extracted atomic relations. Thus, our work on atomic relations extraction that we first described in (Filatova & Hatzivassiloglou 2003) is close to the task of *Unrestricted Relation Discovery* represented by OIE (Shinyama & Sekine 2006) and TextRunner (Banko et al. 2007) systems rather than to the classic MUC-style IE systems.

The necessity of having pre-defined relation types for classic MUC-style IE systems might cause a loss of useful information about the elements which are not anticipated and therefore not considered important by the pre-defined relations. In comparison to classic IE systems described in Section 3.6.2.1, the advantage of atomic relations (and IE systems described in Section 3.6.2.2) is that they have a potential of capturing a variety of relations that are potentially important for the input collection of documents. Thus, in terms of *recall* and *precision*, atomic relations have higher recall than classic IE systems. On the other hand, neither the atomic relation labeling procedure, nor the systems described in Section 3.6.2.2 have a machinery capable of labeling the extracted relations with exact semantic types. Besides, these systems extract many relations that cannot be assigned to any semantic type and are merely noise captured from the input corpus. Thus, atomic relations as well the as systems described in Section 3.6.2.2 have lower precision in comparison to the classic MUC-style IE systems described in Section 3.6.2.1.

In contrast to the OIE (Shinyama & Sekine 2006) and TextRunner (Banko et al. 2007) systems, we do not try to identify all possible relations; rather, given a set of related documents, we attempt to discover the underlying structure of the events described in this document collection. In this respect, atomic relations are close to the information discovered in (Shinyama & Sekine 2006, Sekine 2006). The difference lies in how the discovered relations are used. The systems described in (Shinyama & Sekine 2006, Sekine 2006) group the discovered relations according to the type of information captured by these relations and create tables corresponding to each group of relations. In our research, we focus on sentence selection using the discovered relations for summarization and question-answering (Chapter 5) tasks.

3.7 Formal Model for Information Packaging and Selection

In this section we discuss a general information selection model. We treat information selection as a three-component problem:

1. the identification of the textual units into which the input text should be broken. These text units are later used as building blocks for the final summary;
2. the identification of the textual features which are associated with the important concepts described in the input text;
3. the identification of the appropriate algorithm for selecting the textual units to be included in the final summary.

We focus on the latter two of those steps and explore interdependencies between the choice of features (step 2) and selection algorithm (step 3). As features we used atomic relations described above, and the selection algorithm is described in this section. We experimentally test our hypothesis that event-based (atomic relation) features are helpful for summarization by comparing the performance of three sentence selection algorithms when we use such features versus the case where we use another, widely used set of textual features: the words in the input texts, weighted by their *tf*idf* scores.

By integrating redundancy checking into the selection of the textual units we provide a unified framework for addressing content overlap that does not require external measures of similarity between textual units. We also account for the partial overlap of information between textual units (e.g., a single shared clause), a situation which is common in natural language. However, not many systems have methodologies for dealing with this issue. One of the systems that takes into account the situation of partial overlap of information in textual units and leverages this overlap for creating a more concise summary in described in (Barzilay & McKeown 2005).

Our model for selecting and packing information across multiple text units relies on three components that are specified by each application. First, we assume that there is a finite set T of *textual units* $t_1, t_2, \ldots, t_n$, a subset of which will form the answer or summary. For most approaches to summarization and question answering, which follow the extraction paradigm, the textual units t_i will be obtained by segmenting the input text(s) at an application-specified granularity level, so each t_i would typically be a sentence or paragraph.

Second, we posit the existence of a finite set C of *conceptual units* $c_1, c_2, \ldots, c_m$. The conceptual units encode the information that should be present in the output, and they can be defined in different ways according to the task at hand and the priorities of each system. Obviously, defining the appropriate conceptual units is a core problem, akin to feature selection in machine learning: there is no exact definition of what an important concept is that would apply to all tasks. Current summarization systems often represent concepts indirectly via textual features that give high scores to the textual units that contain important information and should be used in the summary and low scores to those textual units which are not likely to contain information worth including in the final output. In Section 3.1 we provided a description of the features that are currently used by summarization systems.

3.7.1 Full Correspondence

After the sets T and C of textual and conceptual units are formally defined, we can present a mapping function between these textual units and conceptual units. This mapping, a function $f : T \times C \to [0, 1]$, tells us how well each conceptual unit is covered by a given textual unit. Presumably, different approaches will assign different coverage scores for even the same sentences and conceptual units, and the consistency and quality of these scores would be one way to determine the success of each competing approach.

We first examine the case where the function f is limited to zero or one values, i.e., each textual unit either contains/matches a given conceptual feature or not. This is the case with many simple features, such as words and sentence position. Then, we define the total information covered by any given subset S of T (a proposed summary or answer) as

$$I(S) = \sum_{i=1,\ldots,m} w_i \cdot \delta_i \tag{3.1}$$

where w_i is the weight of the concept c_i and

$$\delta_i = \begin{cases} 1, \text{if } \exists j \in \{1, \ldots, m\} \text{ such that } f(t_j, c_i) = 1 \\ 0, \text{otherwise} \end{cases}$$

In other words, the information contained in a summary is the sum of the weights of the conceptual units covered by at least one of the textual units included in the summary.

Of course, the procedure of mapping contextual units onto a set of textual units depends on the nature of the conceptual and textual units. Textual units can be non-sequential. Conceptual units an be identified in such a way that this mapping procedure becomes very complicated. In this work, however, we do not deal with this issue and assume that the mapping procedure is straitforward.

3.7.2 Partial Correspondence between Textual and Conceptual Units

Depending on the nature of the conceptual units, the assumption of a $0-1$ mapping between textual and conceptual units may or may not be practical or even feasible. For example, if the conceptual units represent named entities (a common occurrence in list-type long answers), a partial match between a name found in a text and another name is possible; handling these two names as distinct concepts would be inaccurate. For the experiments presented in this work we assume that a conceptual unit is either completely covered by a text unit or absent in a text unit.

Partial matches between textual and conceptual units introduce a new problem, however: if two textual units partially cover the same concept, it is not apparent to what extent the coverage overlaps. Thus, there are multiple ways to revise equation (3.1) in order to account for partial matches, depending on how conservative we are on the expected overlap. One such way is to assume minimum overlap (the most conservative assumption) and define the total information in the summary as

$$I(S) = \sum_{i=1,\ldots,m} w_i \cdot \max_j f(t_j, c_i) \tag{3.2}$$

Note that this equation reduces to our original formula for information content (equation (3.1)) if the mapping function f only produces 0 and 1 values.

3.7.3 Length and Textual Constraints

We have provided formulae that measure the information covered by a collection of textual units under different mapping constraints. Obviously, we want to maximize this information content. However, this can only sensibly happen when additional constraints on the number or length of the selected textual units are introduced; otherwise, the full set of available textual units would be a solution that proffers a maximal value for equations (3.1)–(3.2), i.e., $\forall S \subset T, I(S) \leq I(T)$. We achieve this by assigning a cost p_i to each textual unit t_i, $i = 1, \ldots, n$, and defining a function P over a set of textual units that provides the total penalty associated with selecting those textual units as the output. In our abstraction, replacing a textual unit with one or more textual units that provide the same content should only affect the penalty, and it makes sense to assign the same cost to a long sentence as to two sentences produced by splitting the original sentence. Also, a shorter sentence should be preferable to a longer sentence with the same information content. Hence, our operational definitions for p_i and P are

$$p_i = \text{length}(t_i), \quad P(S) = \sum_{t_i \in S} p_i \tag{3.3}$$

i.e., the total penalty is equal to the total length of the answer in some basic unit (e.g., words).

Note, however, than in the general case the p_i's need not depend solely on the length, and the total penalty does not need to be a linear combination of them. The cost function can depend on features other then length, for example, number of pronouns—the more pronouns used in a textual unit, the higher the risk of dangling references and the higher the price should be. Finding the best cost function is an interesting research problem by itself.

With the introduction of the cost function $P(S)$ our model has two generally competing components. One approach is to set a limit on $P(S)$ and optimize $I(S)$ while keeping $P(S)$ under that limit. This approach is similar to that taken in evaluations that keep the length of the output summary within certain bounds, such as the recent major summarization evaluations in the Document Understanding Conferences from 2001 to the present (Harman & Voorhees 2001). Another approach would be to combine the two components and assign a composite score to each summary, essentially mandating a specific tradeoff between recall and precision; for example, the total score can be defined as a linear combination of $I(S)$ and $P(S)$, in which case the weights specify the relative importance of coverage and precision/brevity, as well as accounting for scale differences between the two metrics. This approach is similar to the calculation of recall, precision, and F-measure adopted in the recent NIST evaluation of long answers for definition questions (Voorhees 2003*a*). Thus, we will follow the first tactic of maximizing $I(S)$ with a limit on $P(S)$ rather than attempting to solve the thorny issues of weighing the two components appropriately.

3.7.4 Handling Redundancy in Summarization

Identifying redundancy of information has been found useful in determining what text pieces should be included during summarization, on the basis that information that is repeated is likely to be central to the topic or event being discussed. Earlier work has also recognized that, while it is a good idea to select among the passages repeating information, it is also important to avoid repetition of the same information in the final output.

Two main approaches have been proposed for avoiding redundancy in the output. One approach relies on grouping together potential output text units on the basis of their similarity, and producing only a representative from each group (Hatzivassiloglou, Klavans, Holcombe, Barzilay, Kan & McKeown 2001). Sentences can be clustered in this manner according to word overlap, or by using additional content similarity features. This approach has been recently applied to the construction of paragraph-long answers (e.g., (Barzilay & Elhadad 1997, Yu & Hatzivassiloglou 2003, Blair-Goldensohn, McKeown & Schlaikjer 2004)).

An alternative approach, proposed for the synthesis of information during query-based passage retrieval is the maximum marginal relevance (MMR) method (Goldstein et al. 2000). This approach assigns to each potential new sentence in the output a similarity score with the sentences already included in the summary. Only those sentences that contain a substantial amount of *new information* can get into the summary. MMR bases this similarity score on word overlap and additional information about the time when each document was released, and thus can fail to identify repeated information when paraphrasing is used to convey the same meaning.

In contrast to these approaches, our model handles redundancy in the output at the same time it selects the output sentences. It is clear from equations (3.1)–(3.2) that each conceptual unit is counted only once whether it appears in one or multiple textual units. By adding to the final summary text units that contain most important conceptual units Thus, when we find the subset of textual units that maximizes overall information coverage with a constraint on the total number or length of textual units, the model will prefer the collection of textual units that have minimal overlap of covered conceptual units.

3.7.5 Applying the Model

Having presented a formal metric for the information content (and optionally the cost) of any potential summary or answer, the task that remains is to optimize this metric and select the corresponding set of textual units for the final output. As stated in Section 3.7.3, one possible way to do this is to focus on the information content metric and introduce an additional constraint, limiting the total cost to a constant. An alternative is to optimize directly the composite function that combines cost and information content into a single number.

We examine the binary case where each conceptual unit is either present or absent in a textual unit, and the total information content is specified by equation (3.1). The complexity of the problem depends on the cost function, and whether we optimize $I(S)$ while keeping $P(S)$ fixed or whether we optimize a combined function of both of those quantities. We will only consider the former case in the present work. We start by examining an artificially simple case, where the cost assigned to each textual unit is 1, and the function P for combining costs is their sum. In this case, the total cost is equal to the number of textual units used in a summary.

This problem, as we have formalized above, is identical to the *Maximum Set Coverage* problem studied in theoretical computer science: given C, a finite set of weighted elements, a collection T of subsets of C, and an integer k, find those k sets that maximize the total number of elements in the union of T's members (Hochbaum 1997). In our case, the zero-one mapping allows us to view each textual unit as a subset of the conceptual units space, containing those conceptual units covered by the textual unit, and k is the total target cost. Unfortunately, *maximum set coverage* is NP-hard, as it is reducible to the classic *set cover* problem (given a finite set and a collection of subsets of that set, find the smallest subset of that collection whose members' union is equal to the original set) (Hochbaum 1997). It follows that more general formulations of the cost function that actually are more realistic for our problem (such as defining the total cost as the sum of the lengths of the selected textual units and allowing the textual units to have different lengths) will also result in an NP-hard problem, as we can reduce these versions to the special case of *maximum set coverage*.

Nevertheless, the correspondence with maximum set coverage provides a silver lining. Since the problem is known to be NP-hard, properties of simple greedy algorithms have been explored, and a straightforward local maximization method has been proved to give solutions within a known bound of the optimal solution. The greedy algorithm for maximum set coverage is: Start with an empty solution S, and iteratively add to the S the set T_i that maximizes $I(S \cup T_i)$. It is provable that this algorithm is the best polynomial approximation algorithm for the problem (Hochbaum 1997), and that it achieves a solution bounded as follows

$$I(\textsc{Opt}) \geq I(\textsc{Greedy}) \geq \left[1 - \left(1 - \frac{1}{k}\right)^k\right] I(\textsc{Opt})$$

$$> \left(1 - \frac{1}{e}\right) I(\textsc{Opt}) \approx 0.6321 \times I(\textsc{Opt})$$

where $I(\textsc{Opt})$ is the information content of the optimal summary and $I(\textsc{Greedy})$ is the information content of the summary produced by this greedy algorithm.

For the more realistic case where cost is specified as the total length of the summary, and where we try to optimize $I(S)$ with a limit on $P(S)$ (see Section 3.7.3), we propose two greedy algorithms inspired by the algorithm above. Both our algorithms operate by first calculating a ranking of the textual units in decreasing order. This ranking is for the first algorithm, which we call *adaptive greedy algorithm*, identical to the ranking provided by the basic greedy algorithm, i.e., each textual unit receives as score the increase in $I(S)$ that it generates when added to the output, in the order specified by the basic greedy algorithm. Our second greedy algorithm (dubbed *modified greedy algorithm* below) modifies this ranking by prioritizing the conceptual units with highest individual weight w_i; it ranks first the textual unit that has the highest contribution to $I(S)$ while covering this conceptual unit with the highest individual weight, and then iteratively proceeds with the textual unit that has the highest contribution to $I(S)$ while covering the next most important unaccounted for conceptual unit.

Given the rankings of textual units, we can then produce an output of a given length by adopting appropriate stopping criteria for when to stop adding textual units (in order according to their ranking) to the output. There is no clear rule for conforming to a specific length (for example, DUC 2001 allowed submitted summaries to go over “a reasonable percentage” of the target length, while DUC 2004 cuts summaries mid-sentence at exactly the target length). As the summary length in DUC is measured in words, in our experiments we extracted the specified number of words out of the top sentences (truncating the last sentence if necessary).

3.8 Experiments

We chose as our input data the document sets used in the evaluation of multidocument summarization during the first Document Understanding Conference (DUC), organized by NIST (Harman & Voorhees 2001). This collection contains 30 test document sets, each with approximately 10 news stories on different events; document sets vary significantly in their internal coherence. For each document set three human-constructed summaries are provided for each of the target lengths of 50, 100, 200, and 400 words. We selected DUC 2001 because ideal summaries are available for multiple lengths.

Conceptual and Textual Units Our textual units are sentences, while the features representing concepts are either atomic relations, as described in Section 3.3, or a fairly basic and widely used set of lexical features, namely the list of words present in each input text. The algorithm for extracting atomic relation triplets assigns a weight to each such triplet, while for words we used as weights their *tf*idf* values.[9]

Evaluation Metric Given the difficulties in coming up with a universally accepted evaluation measure for summarization, and the fact that obtaining judgments by humans is time-consuming and labor-intensive, we adopted an automated process for comparing system-produced summaries to “ideal” summaries written by humans. The method, ROUGE (Lin

[9] *idf* values are taken from `http://elib.cs.berkeley.edu/docfreq/`

& Hovy 2003), is based on n-gram overlap between the system-produced and ideal summaries. As such, it is a recall-based measure, and it requires that the length of the summaries be controlled to allow meaningful comparisons. For the evaluation described in this thesis we used version ROUGE-1.5.5.

ROUGE can be readily applied to compare the performance of different systems on the same set of documents, assuming that ideal summaries are available for those documents. At the same time, ROUGE scores are difficult to interpret as they are not absolute and are not comparable across source document sets. Nevertheless, ROUGE is widely used for summarization evaluation as this is the only absolutely automatic evaluation system currently available.

In our comparison, we used as reference summaries those created by NIST assessors for the DUC task of generic summarization. The human annotators may not have created the same models if asked for summaries describing the major events in the input texts instead of generic summaries.

In our experiments, we used two ROUGE metrics for comparing the systems based on two sets of features: the metric based on calculating the intersection of unigrams in the model and system-created summaries; and the metric based on the analysis of the longest common substrings that appear in the model and system-created summaries.

Summary Length For a given set of features and selection algorithm we got a sorted list of sentences extracted according to that particular algorithm. Then, for each DUC document set we created summaries of length 50, 100, 200, and 400. The chances that the created summaries have sentence breaks after exactly 50, 100, 200, and 400 words are very low; thus, for each summary we tried three methods for choosing when to stop including sentences in the summary:

1. **Exact**: extract the exact amount of words (50, 100, 200, or 400) out of the top sentences (truncating the last sentence if necessary);

2. **Plus**: extract as many complete sentences as possible until the overall length is at least as much as the target length (50, 100, 200, or 400 words). In constant to the previous **exact** case, we do not truncate the last sentence in the summary;

3. **Minus**: extract as many complete sentences as possible without exceeding the target length. Thus, each summary consists of only complete sentences and its length is less or equal to 50, 100, 200, and 400 words.

We were interested in testing the difference in results for the above three approaches as human models consist of complete sentence only. According to our results, the deviation in result scores for these three truncating methods is insignificant.

Algorithms In our experiments we used three variations of the greedy algorithm for textual unit selection: static, adaptive and modified adaptive algorithms. Thus, for each document set we create 36 summaries taking into consideration:

1. Three different textual unit selection algorithms (static, adaptive, and modified adaptive algorithms);
2. Four different length requirements (50, 100, 200, and 400 words)
3. Three different methods to truncate the last sentence in each summary (exact, plus, and minus).

According to the results described in Sections 3.8.1–3.8.3, on average, the system based on atomic relations outperforms the system based on *tf*idf* features. While Sections 3.8.1–3.8.3 have a detailed overview of the results, Table 3.9 shows the values of significance t-test with confidence level .95. The t-test values were calculated according to the ROUGE scores for each of the 36 summary sets where each set had a summary based on atomic relations and a summary based on *tf*idf* features. According to Table 3.9, only in one case there is no statistical significance in the results (boldfaced number).

3.8.1 Results: Static Greedy Algorithm

In our first experiment, we used the static greedy algorithm to create summaries of various lengths. This algorithm does not support any mechanism for avoiding redundant information in the summary. Instead, it rates each textual unit independently. Textual units are included in the summary if and only if they cover lots of concepts. More specifically,

Mode	t-test (ROUGE–1 Average_F)	t-test (ROUGE–L Average_F)
Static (50 exact)	0.0038123254	0.0029271759
Static (100 exact)	0.0135204640	0.0027712812
Static (200 exact)	0.0000709862	0.0000014668
Static (400 exact)	0.0006905387	0.0000124622
Static (50 plus)	**0.0504490459**	0.0250261587
Static (100 plus)	0.0229099077	0.0046510142
Static (200 plus)	0.0001476399	0.0000212257
Static (400 plus)	0.0024629405	0.0001419934
Static (50 minus)	0.0232754647	0.0233503236
Static (100 minus)	0.0006299687	0.0002024170
Static (200 minus)	0.0000168331	0.0000011855
Static (400 minus)	0.0000441877	0.0000441877
Adaptive (50 exact)	0.0000002141	0.0000001715
Adaptive (100 exact)	0.0000000077	0.0000000226
Adaptive (200 exact)	0.0000073361	0.0000071865
Adaptive (400 exact)	0.0051682329	0.0020621088
Adaptive (50 plus)	0.0000141647	0.0000012860
Adaptive (100 plus)	0.0000075310	0.0000014977
Adaptive (200 plus)	0.0000689853	0.0000054697
Adaptive (400 plus)	0.0098296600	0.0017471479
Adaptive (50 minus)	0.0001168278	0.0003025800
Adaptive (100 minus)	0.0000000839	0.0000000507
Adaptive (200 minus)	0.0000001079	0.0000000990
Adaptive (400 minus)	0.0007199788	0.0004647853
Modified adaptive (50 exact)	0.0000001583	0.0000008542
Modified adaptive (100 exact)	0.0000000571	0.0000000057
Modified adaptive (200 exact)	0.0000000565	0.0000001069
Modified adaptive (400 exact)	0.0001261883	0.0000384846
Modified adaptive (50 plus)	0.0000020777	0.0000006610
Modified adaptive (100 plus)	0.0000101263	0.0000009607
Modified adaptive (200 plus)	0.0000003754	0.0000009064
Modified adaptive (400 plus)	0.0003411072	0.0001142395
Modified adaptive (50 minus)	0.0001846389	0.0004141069
Modified adaptive (100 minus)	0.0000000815	0.0000000079
Modified adaptive (200 minus)	0.0000000014	0.0000000003
Modified adaptive (400 minus)	0.0000054631	0.0000030883

Table 3.9: T-test values for all the text unit selection modes, and summary lengths.

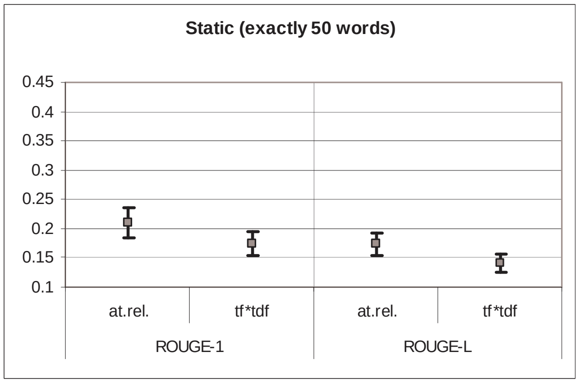


Figure 3.4: Average ROUGE-1 and ROUGE-L scores over 30 DUC 2001 documents (static algorithm, exactly 50 words).

1. For every textual unit, calculate the weight of this textual unit as the sum of the weights of all the concepts covered by this textual unit.

2. Choose the textual unit with the maximum weight and add it to the final output.

3. Continue extracting other textual units in order of total weight till we get the summary of the desired length.

Figures 3.4–3.7 show average ROUGE-1 and ROUGE-L scores (over 30 DUC document sets) for the summaries of exactly 50, 100, 200, and 400 words created using static greedy algorithm. ROUGE-1 scores are based on unigrams, and ROUGE-L scores are based on the longest common subsequence was included in the summary. We chose these two metrics as ROUGE-1 scores have the highest correlation with human scores (Lin & Hovy 2003) and ROUGE-L aims to overcome some deficiencies of ROUGE-1, such as its susceptibility to ungrammatical keyword packing by dishonest summarizers.[10] According to Figures 3.4–3.7, the system based on atomic relations on average outperforms the system based on *tf*idf* features. The advantage, though, is statistically significant for both ROUGE-1 and ROUGE-L scores only for the summaries of length 200.

[10] More detail on the ROUGE evaluation metrics can be obtained online from http://www.isi.edu/4 cyl/papers/ROUGEWorking-Note-v1.3.1.pdf.

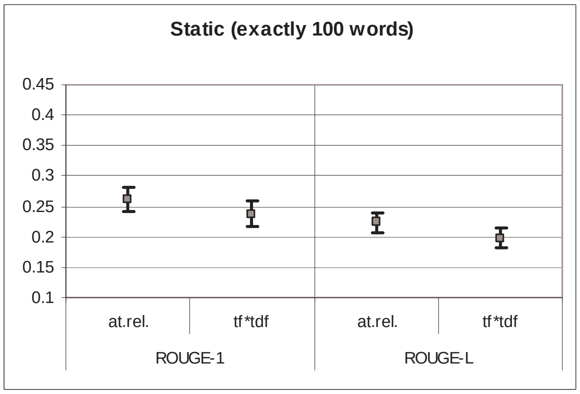


Figure 3.5: Average ROUGE-1 and ROUGE-L scores over 30 DUC 2001 documents (static algorithm, exactly 100 words).

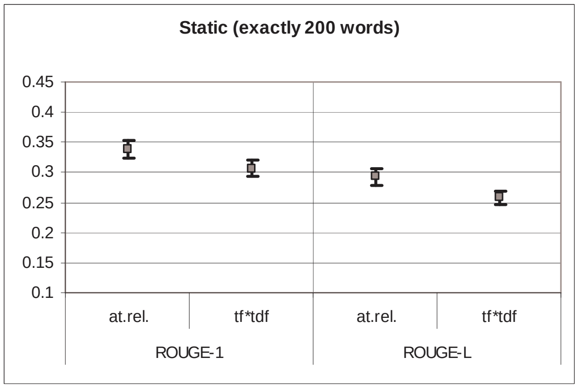


Figure 3.6: Average ROUGE-1 and ROUGE-L scores over 30 DUC 2001 documents (static algorithm, exactly 200 words).

Figures 3.8–3.11 show ROUGE-1 scores (over 30 DUC document sets) for the summaries of exactly 50, 100, 200, and 400 words. According to these figures, for many document sets the performance of the both systems is rather close, but for several document sets the system based on atomic relations outperforms the system based on *tf*idf*. This observation is also supported by the numbers presented in Table 3.9. The difference in the performance of the two systems becomes more profound when the algorithms for textual unit selection that tackle the redundancy problem are applied (Sections 3.8.2–3.8.3).

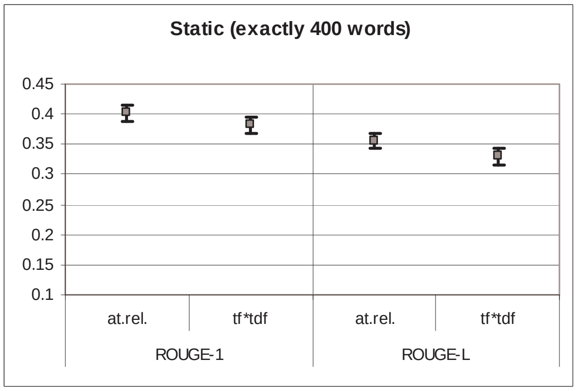


Figure 3.7: Average ROUGE-1 and ROUGE-L scores over 30 DUC 2001 documents (static algorithm, exactly 400 words).

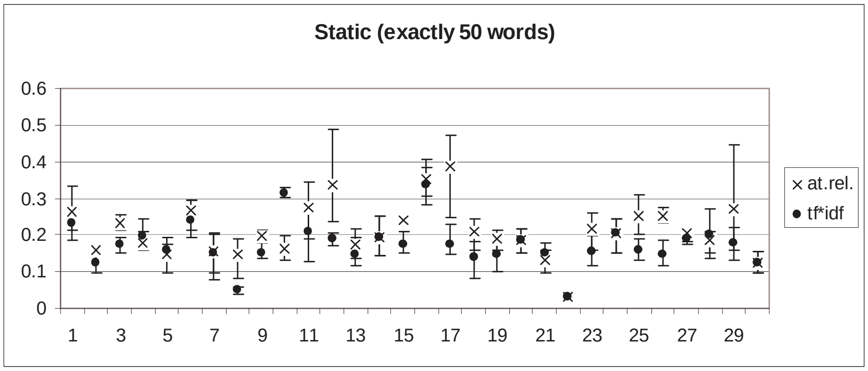


Figure 3.8: ROUGE-1 scores over 30 DUC 2001 documents (static algorithm, exactly 50 words).

3.8.2 Results: Adaptive Greedy Algorithm

For the second experiment we used the adaptive greedy algorithm, which accounts for information overlap across sentences in the summary.

1. For each textual unit calculate its weight as the sum of weights of all concepts it covers.

2. Choose the textual unit with the maximum weight and add it to the output. Add the concepts covered by this textual unit to the list of concepts covered in the final output.

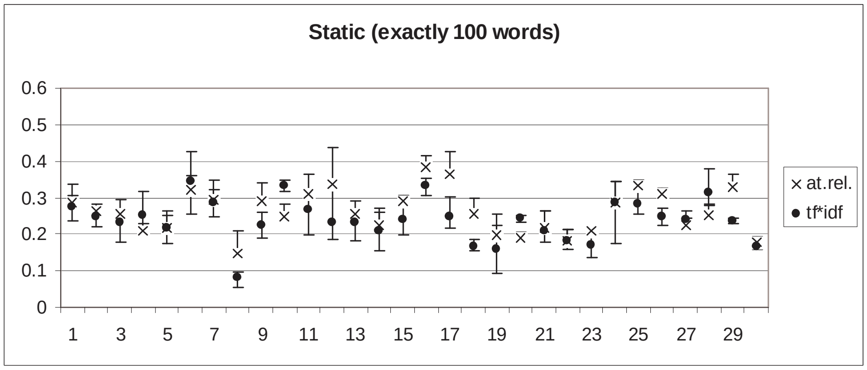


Figure 3.9: ROUGE-1 scores over 30 DUC 2001 documents (static algorithm, exactly 100 words).

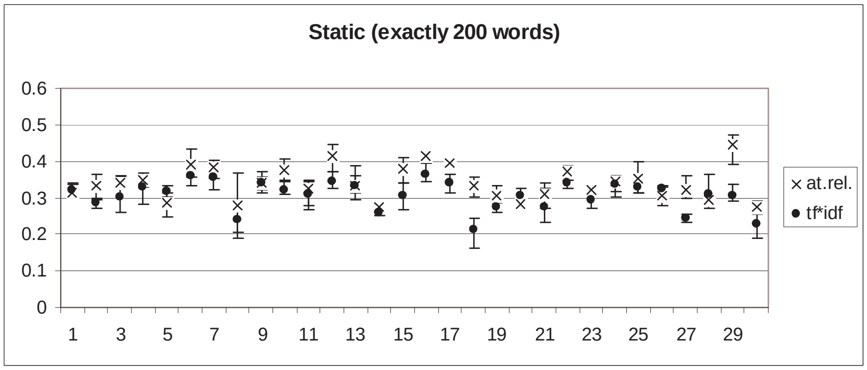


Figure 3.10: ROUGE-1 scores over 30 DUC 2001 documents (static algorithm, exactly 200 words).

3. Recalculate the weights of the textual units: subtract from each unit's weight the weight of all concepts in it that are already covered in the output.

4. Continue extracting text units in order of their total weight (going back to step 2) until the summary is of the desired length.

Figures 3.12–3.15 show average ROUGE-1 and ROUGE-L scores (over 30 DUC document sets) for the summaries of exactly 50, 100, 200, and 400 words. According to these figures, the system based on atomic relations on average outperforms the system based on *tf*idf* features. In contrast to the data presented in Figures 3.4–3.7, the difference in the performance of the two systems is statistically significant for both ROUGE-1 and ROUGE-L

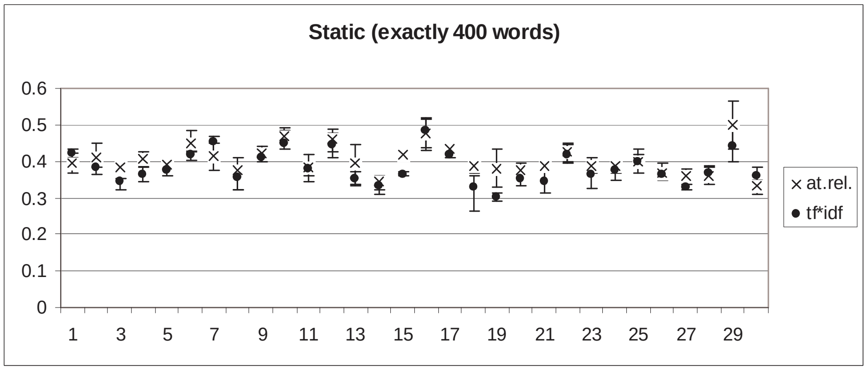


Figure 3.11: ROUGE-1 scores over 30 DUC 2001 documents (static algorithm, exactly 400 words).

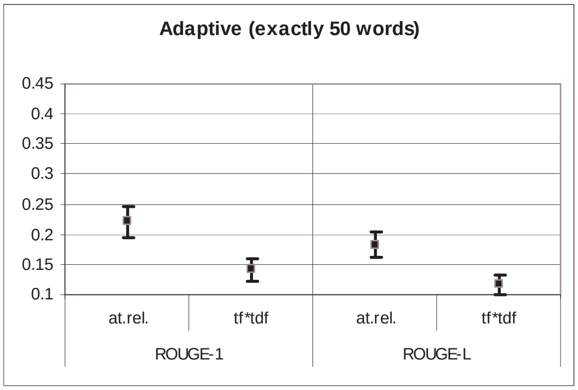


Figure 3.12: Average ROUGE-1 and ROUGE-L scores over 30 DUC 2001 documents (adaptive algorithm, exactly 50 words).

scores for the summaries of all, except 400 word, lengths.

Figures 3.16–3.19 show ROUGE-1 scores (over 30 DUC document sets) for the summaries of exactly 50, 100, 200, and 400 words. According to these figures, in most cases the performance of the atomic relation-based summarizer is higher than the performance of the summarizer based on words and their *tf*idf* values. For some document sets, though, the summarizer based on words and their *tf*idf* values gives better scores. This phenomenon can be explained through an additional analysis of document sets according to their internal coherence. Atomic relation extraction works best for a collection of documents with well-defined constituent parts of events and where documents are clustered around one specific

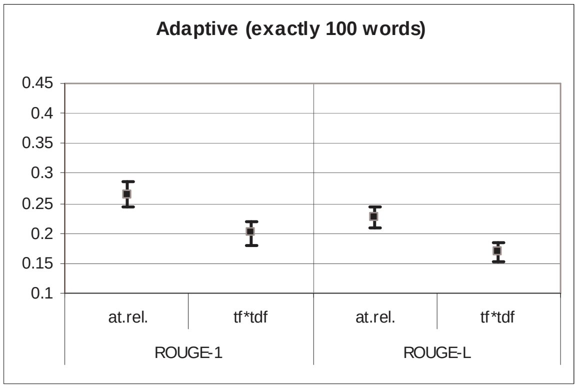


Figure 3.13: Average ROUGE-1 and ROUGE-L scores over 30 DUC 2001 documents (adaptive algorithm, exactly 100 words).

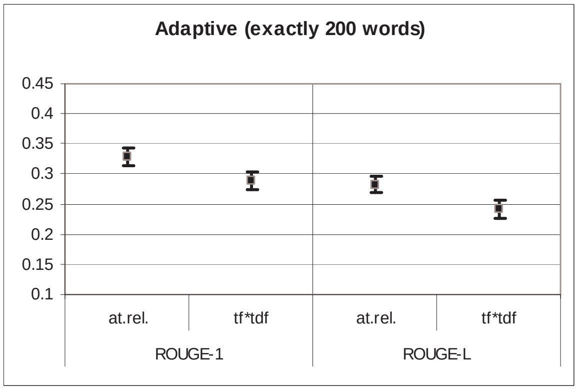


Figure 3.14: Average ROUGE-1 and ROUGE-L scores over 30 DUC 2001 documents (adaptive algorithm, exactly 200 words).

major event. For such document sets atomic relations are good features for basing the summary on. However, some DUC 2001 document sets describe a succession of multiple events linked in time or of different events of the same type (e.g., Clarence Thomas' ascendancy to the Supreme Court, document set 7 in Figures 3.16–3.19) In such cases, a lot of different participants are mentioned with only few common elements (e.g., Clarence Thomas himself). Thus, most of the atomic relations have similar low weights and it is difficult to identify those atomic events that can point out the most important textual units.

The fact that the modification of the simple static greedy algorithm influences more the performance of the system based on atomic relations than the performance of the system

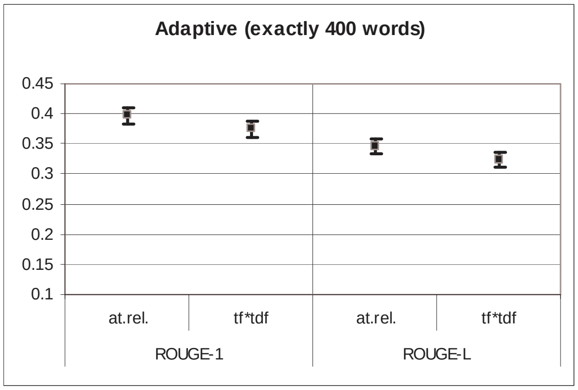


Figure 3.15: Average ROUGE-1 and ROUGE-L scores over 30 DUC 2001 documents (adaptive algorithm, exactly 400 words).

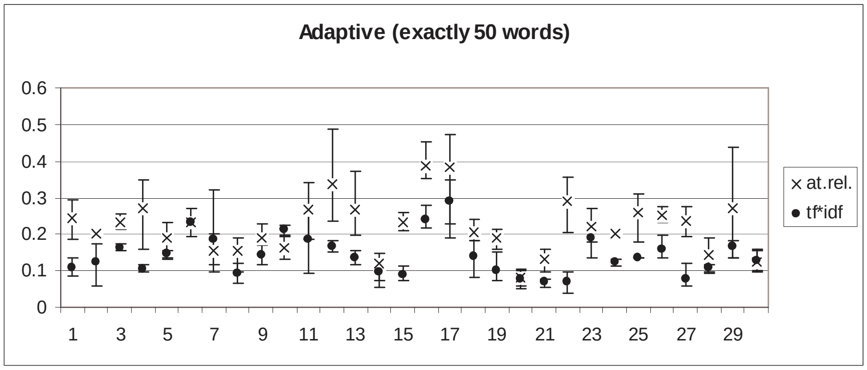


Figure 3.16: ROUGE-1 scores over 30 DUC 2001 documents (adaptive algorithm, exactly 50 words).

based on *tf*idf* scores indicates that using stand alone words as information selection features is not compatible with our information redundancy component. A likely explanation is that words are correlated, and the presence of an important word makes other words in the same sentence also potentially important, a fact not captured by the *tf*idf* feature. Atomic relations, on the other hand, exhibit less of a dependence on each other, since each triplet captures a specific interaction between two entities.

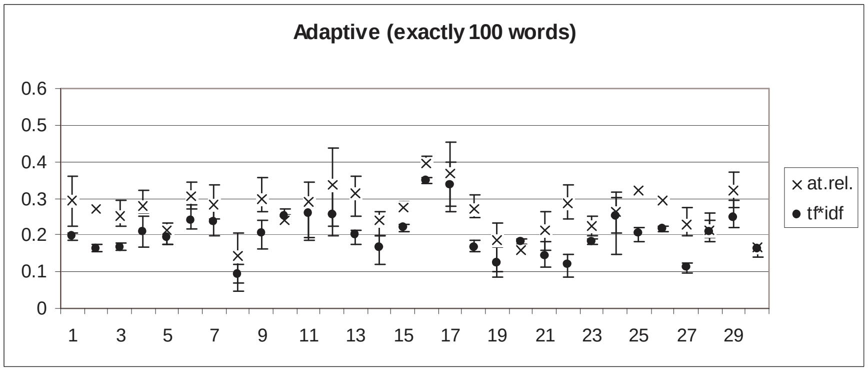


Figure 3.17: ROUGE-1 scores over 30 DUC 2001 documents (adaptive algorithm, exactly 100 words).

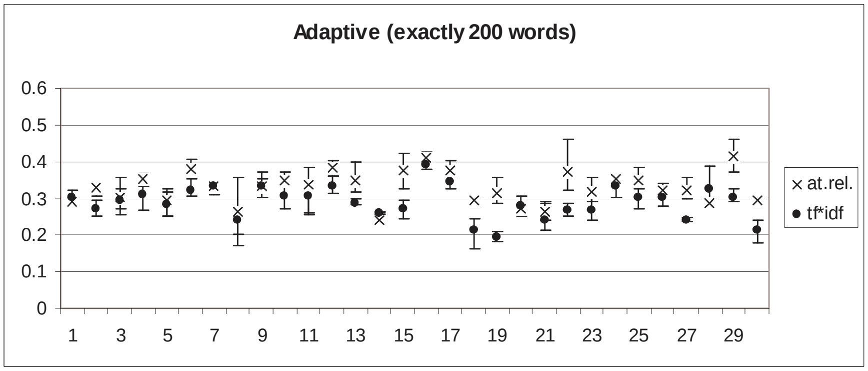


Figure 3.18: ROUGE-1 scores over 30 DUC 2001 documents (adaptive algorithm, exactly 200 words).

3.8.3 Results: Modified Adaptive Greedy Algorithm

The last approach for selecting text units that we tried was Modified Adaptive Greedy Algorithm.

1. For every textual unit calculate its weight as the sum of weights of all concepts it covers.

2. Consider only those textual units that contain the concept with the highest weight that has not yet been covered. Out of these, choose the one with highest total weight and add it to the final output. Add the concepts which are covered by this textual unit to the list of concepts covered in the final output.

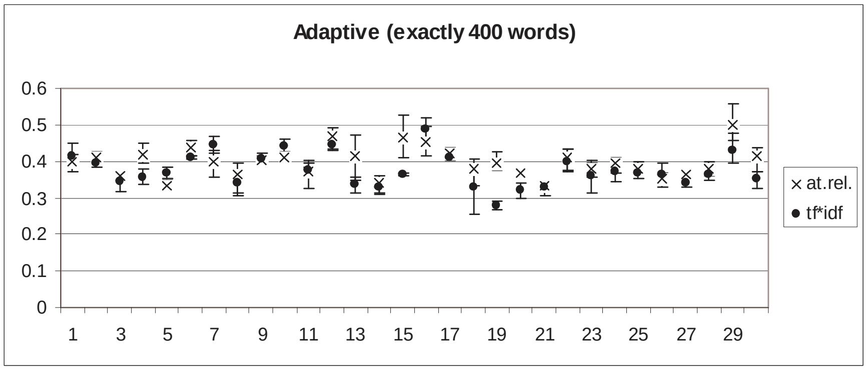


Figure 3.19: ROUGE-1 scores over 30 DUC 2001 documents (adaptive algorithm, exactly 400 words).

3. Recalculate the weights of the textual units: subtract from each unit's weight the weight of all concepts in it that are already covered in the output.

4. Continue extracting textual units, going back to step 2 each time, until we get a summary of the desired length.

Figures 3.20–3.23 show average ROUGE-1 and ROUGE-L scores (over 30 DUC document sets) for the summaries of exactly 50, 100, 200, and 400 words. According to these figures, the performance of the system based on atomic relations, on average, is significantly better than the performance of the system based on words and their *tf*idf* scores. For the modified adaptive greedy algorithm this is true for the summaries of all lengths (50, 100, 200, and 400 words) for both ROUGE-1 and ROUGE-L scores. In other words, the prioritization of individual important concepts addresses the correlation between words and allows the summarizer to benefit from redundancy reduction.

Figures 3.24–3.27 show ROUGE-1 scores (over 30 DUC document sets) for the summaries of exactly 50, 100, 200, and 400 words.

3.9 Results Discussion

Our experimental results show that atomic relations can be successfully used as an approximation of conceptual units for summarization. We also experimentally confirmed that

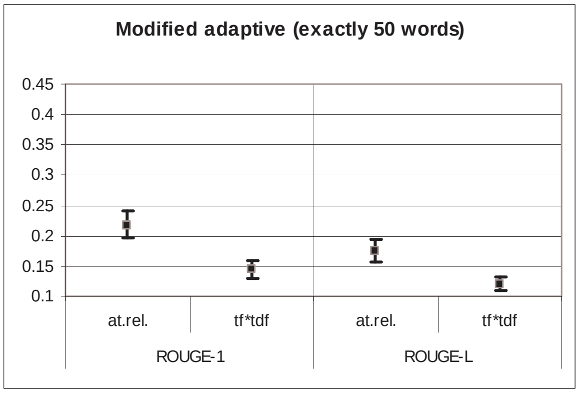


Figure 3.20: Average ROUGE-1 and ROUGE-L scores over 30 DUC 2001 documents (modified

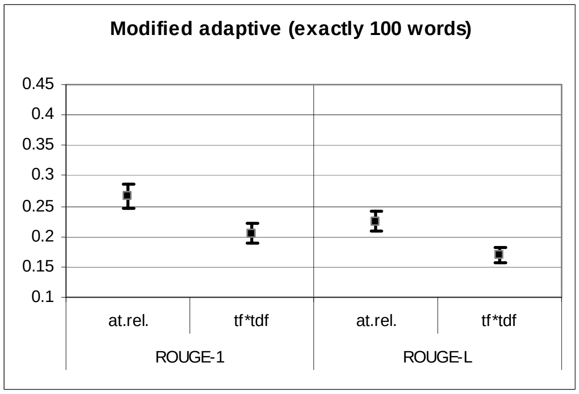


Figure 3.21: Average ROUGE-1 and ROUGE-L scores over 30 DUC 2001 documents (modified adaptive algorithm, exactly 100 words).

addressing redundancy elimination is a crucial step in improving the summary quality. The boost in the performance of atomic relation-based summarization system was significant when we used the modified static greedy algorithm and lowered the scores of the sentences that contain relations already present in the summary. We also showed that this improvement is more profound for the atomic relation-based summarizer than for the summarizer based on words and their *tf*idf* scores. This provides evidence that the information selection process guided by connections among important words and concepts is a more promising direction than the information selection process guided by stand alone words.

It must be noted that the system based the shallow semantic network works best for the

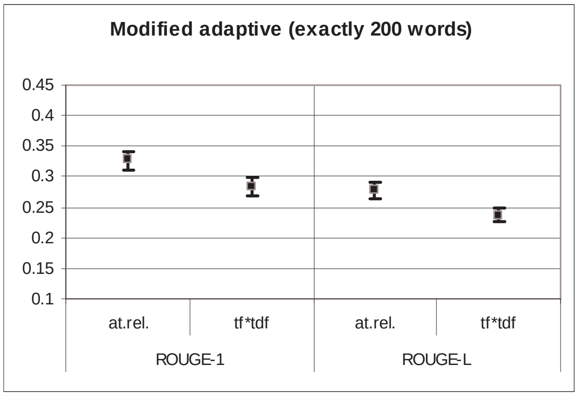


Figure 3.22: Average ROUGE-1 and ROUGE-L scores over 30 DUC 2001 documents (modified adaptive algorithm, exactly 200 words).

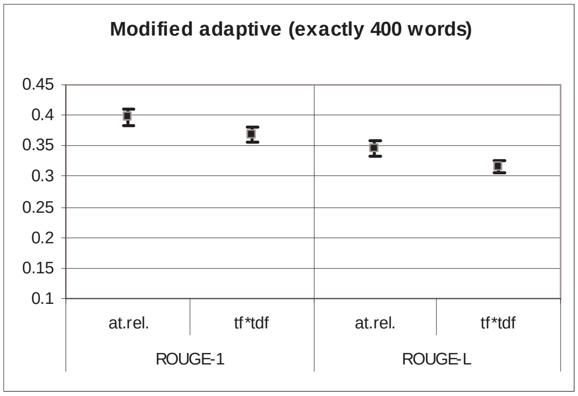


Figure 3.23: Average ROUGE-1 and ROUGE-L scores over 30 DUC 2001 documents (modified adaptive algorithm, exactly 400 words).

document collection that cover events with a well-defined set of constituent parts (participants, locations, dates) within a rather short time period. For the document collections that do not satisfy this criteria, the system based on words and their *tf*idf* scores outperforms the system based on a shallow semantic network and the system. This situation is typical for the DUC task. No system performed consistently better than the rest of the systems on all the document collections. All summarization system (including DUC systems) are trained for a specific document collection type, or domain. There does not exist a single generic summarization system that handles well all types of input document collections.

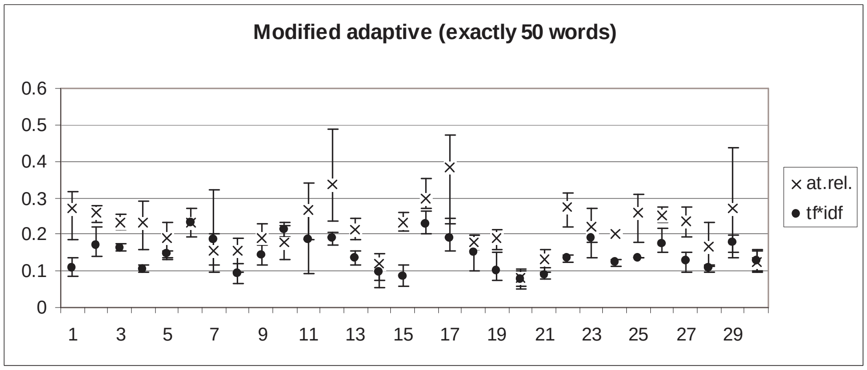


Figure 3.24: ROUGE-1 scores over 30 DUC 2001 documents (modified adaptive algorithm, exactly 50 words).

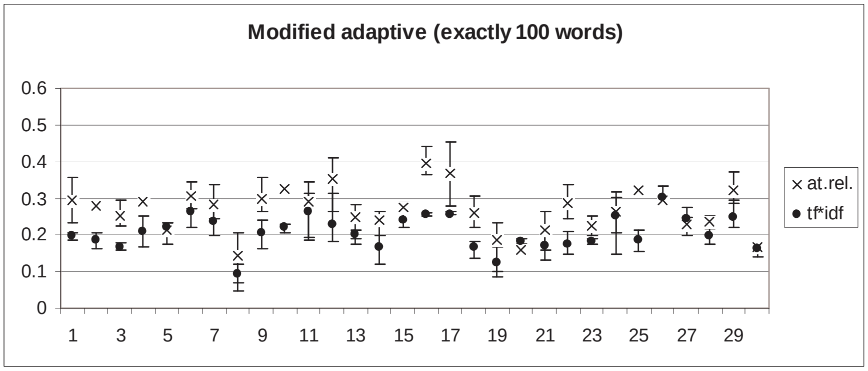


Figure 3.25: ROUGE-1 scores over 30 DUC 2001 documents (modified adaptive algorithm, exactly 100 words).

3.9.1 Conclusion

In this chapter, we introduced atomic relations as a feature that can be automatically extracted from text and used for summarization. These relations connect the entities that are described in a collection of documents among each other. Atomic relations are most effective when analyzed as a collective representation of a set of documents, rather than as stand-alone relations. We call this collective representation a *shallow semantic network*. We presented experimental evidence that the relation-based representation of documents was superior to the techniques that relied on term frequencies.

Another contribution described in this chapter is a formal model of information pack-

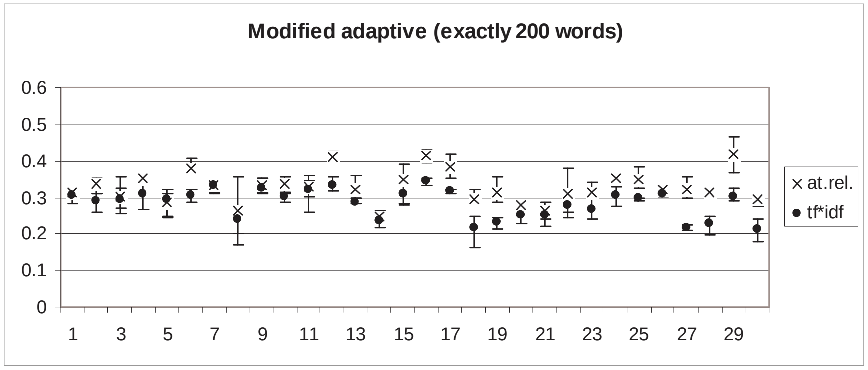


Figure 3.26: ROUGE-1 scores over 30 DUC 2001 documents (modified adaptive algorithm, exactly 200 words).

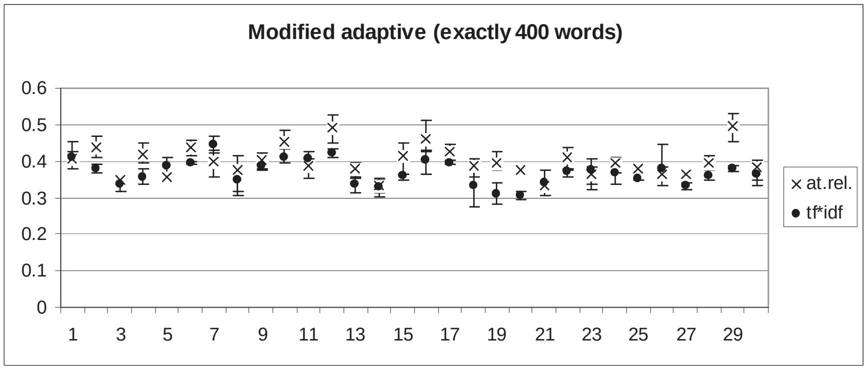


Figure 3.27: ROUGE-1 scores over 30 DUC 2001 documents (modified adaptive algorithm, exactly 400 words).

aging and selection that utilizes this feature to select sentences for the summary while minimizing the overlap of information in the output. Our experimental results indicate that atomic relations are indeed an effective feature, at least in comparison with words in the input texts that form the basis of many of current summarizers' feature sets. With all three of our summarization algorithms, we achieved a gain in performance when using events. This gain was actually more pronounced with the more sophisticated sentence selection methods, establishing that events also exhibit less interdependence than features based directly on words. The advantage was also larger in longer summaries.

Chapter 4

Domain Modelling

Happy families are all alike;
every unhappy family is
unhappy in its own way.

"Anna Karenina"
Leo Tolstoy (translation by
Constance Garnett)

Open-ended question-answering (QA) systems typically produce a response containing a variety of specific facts proscribed by the question type. A biography, for example, might contain the date of birth, occupation, or nationality of the person in question (Duboue & McKeown 2003, Zhou et al. 2004, Weischedel, Xu & Licuanan 2004, Filatova & Prager 2005). A definition may contain the genus of the term and characteristic attributes (Blair-Goldensohn, McKeown & Schlaikjer 2004). A response to a question about a terrorist attack might include the event name, location, victims, perpetrator and date as the templates designed for the Message Understanding Conferences predicted (Radev & McKeown 1998, White et al. 2001). Furthermore, the type of information included varies depending on context. A biography of an actor would include movie names, while a biography of an inventor would include the names of inventions. A description of a terrorist event in Latin America in the eighties is different from the description of today's terrorist events.

How does one determine what facts are important for different kinds of responses? Often the types of facts that are important are hand encoded ahead of time by a human expert (e.g., as in the case of MUC templates). In this chapter, we present an approach that allows a system to learn the types of facts that are appropriate for a particular response. We focus on acquiring fact-types for events, automatically producing a *template* that can guide the creation of responses to questions requiring a description of an event. The template can be tailored to a specific time period or country simply by changing the document collections from which learning takes place.

In this work, a *domain* is a set of events of a particular type; ***earthquake*** and ***presidential election*** are two such domains. Domains can be instantiated by several *instances* of events of that type (e.g., the earthquake in Japan in October 2004, the earthquake in Afghanistan in March 2002, etc.).[1] The granularity of domains and instances can be altered by examining data at different levels of detail, and domains can be hierarchically structured. An ideal template is a set of attribute-value pairs, with the attributes specifying particular functional roles important for the domain events.

In this chapter we present a novel method for automatic induction of domain templates. Most of the currently used domain templates are created manually; this manual procedure is time-consuming and the created domain templates are not portable across different domains. Our method of domain-independent on-the-fly template induction is completely automatic. As input it requires several document collections describing domain instances. We cross-examine the input instances, we identify verbs important for the majority of instances and relationships containing these verbs. We generalize across multiple domain instances to automatically determine which of these relations should be used in the template. We report on data collection efforts and results from four domains. We assess how well the automatically produced templates satisfy users' needs, as manifested by questions collected for these domains.

[1] As it has been noted in Chapter 2, NLP terminology is not standardized across different tasks. Two NLP tasks most close to our research are Topic Detection and Tracking (TDT) (Fiscus et al. 1999) and Information Extraction (IE) (Marsh & Perzanowski 1997). In TDT terminology, our domains are topics and our instances are events. In IE terminology, our domains are scenarios and our domain templates are scenario templates.

4.1 Related Work

```
Scene 1: ENTERING
              S PTRANS S into restaurant
              S ATTEND eyes to tables
              S MBUILD where to sit
              S PTRANS S to table
              S MOVE S to sitting position

Scene 2: ORDERING
              S PTRANS menu to S (menu already on table)
              S MBUILD choice of food
              S MTRANS signal to waiter
              waiter PTRANS to table
              S MTRANS 'I want food' to waiter
              waiter PTRANS to cook

Scene 3: EATING
              Cook ATRANS food to waiter
              waiter PTRANS food to S
              S INGEST food

Scene 4: EXITING
              waiter MOVE write check
              waiter PTRANS to S
              waiter ATRANS check to S
              S ATRANS money to waiter
              S PTRANS out of restaurant
```

Figure 4.1: Restaurant visit script (S - customer).

In the middle of the 1990s, domain templates were suggested as a means of encoding world knowledge about a particular domain. Deciding what slots to include in the template, and what restrictions to place on their potential fillers, is a knowledge representation problem (Hobbs & Israel 1994). Templates were used in the main IE competitions, the Message Understanding Conferences (Hobbs & Israel 1994, Onyshkevych 1993, Marsh & Perzanowski 1997). One of the recent evaluations, ACE,[2] uses pre-defined frames connecting event types (e.g., *arrest*, *release*) to a set of attributes. For example, ACE ATTACK events have three participant slots (ATTACKER-ARG, TARGET-ARG, and INSTRUMENT-ARG) and two attribute slots (TIME-ARG and PLACE-ARG).[3] Thus, ACE definitions of events and relations are very similar to MUC templates. The template *construction* task is not addressed by classic IE systems (e.g., MUC and ACE information extraction systems). The domain templates were created manually by experts to capture the structure of the facts sought. Our system automatically generates a template that captures the generally most important information for a particular domain and is reusable across multiple instances of that domain.

To fill in MUC templates IE systems used a variety of structures similar to Case Frames suggested in (Lehnert, Cardie, Fisher, McCarthy, Riloff & Soderlan 1994). Examples of manually created case frames for several verbs identified as important for a *terrorist attack* domain are presented in Figure 4.2. It must be noted that both verbs for which case frames are created as well as the semantic arguments for these verbs are identified manually.

The work closest to ours is the one on Conceptual Case Frame Acquisition (Riloff & Schmelzenbach 1998). This work attempts to create the above case frames automatically. It uses patterns automatically acquired by the AutoSlog-TS system (Riloff 1996) and a semantic lexicon created by the system described in (Riloff & Shepherd 1997). As input, AutoSlog-TS requires only a set of manually defined syntactic patterns and preclassified texts. In our system, we do not use a predefined set of syntactic patterns, instead, we rely on co-occurrences of elements in syntactic trees. After the list of extraction patterns is constructed, a human annotator refines this list by removing irrelevant patterns. Another required set of data for automatic case frame acquisition procedure is a semantic lexicon. The procedure for semantic lexicon acquisition described in (Riloff & Shepherd 1997) uses a

[2] `http://www.nist.gov/speech/tests/ace/index.htm`

[3] `http://projects.ldc.upenn.edu/ace/docs/English-Events-Guidelines_v5.4.3.pdf`

ATTACK (passive-verb “attacked”)	
Victim	= subject
Target	= subject
Perpetrator	= pp(by)
Instrument	= pp(by)
ACCUSATION (active-verb “blamed”)	
Accuser	= subject
Perpetrator	= direct object
Perpetrator	= pp(on)
SABOTAGE (noun “sabotage”)	
Perpetrator	= pp(by)
Instrument	= pp(with)
Location	= pp(on)
Victim	= pp(against), pp(of), pp(on)
Target	= pp(against), pp(of), pp(on)

Figure 4.2: Examples of manually created case frames for the *terrorist attack* domain.

bootstrapping approach. As input, this bootstrapping approach requires a set of semantic classes for which lexicons are acquired; plus, each semantic class requires a small set of *seed lexicon elements* to start the bootstrapping procedure. Both domain semantic classes and seed lexicon elements are created manually. After the semantic lexicon is constructed, a human annotator reviews it and removes irrelevant elements. The presence of a human annotator in the Case Frame acquisition loop is an important requirement. The annotator who filters the sets of extraction patterns and dictionaries, and who creates lists of elements associated with the pre-defined semantic classes should have knowledge of the domain under investigation. In our system, we rely on named entity tags and nouns that frequently co-occur with verbs that are automatically identified as important for a domain. Thus, our approach does not require any domain-specific knowledge and uses only corpus-based statistics. The advantages and limitations of our completely automatic approach are discussed in Section 4.5. The rest of this section describes other research close to automatic domain template induction such as, work on GISTexter summarization, Topic Themes, atomic relations, etc.

The GISTexter summarization system (Harabagiu & Maiorano 2002) used statistics over an arbitrary document collection together with semantic relations from WordNet. The created templates heavily depend on the topical relations encoded in WordNet. In many newswire articles the important topical relations include proper names, which are not mentioned in WordNet. In our work, we learn templates from several collections of documents aiming for a general domain template. We rely on the relations which are cross-mentioned in different instances of the domain rather then on WordNet topic relations used in one of these instances. In our work, we learn domain templates by cross-examining several collections of documents on the same topic, aiming for a general domain template. We rely on relations cross-mentioned in different instances of the domain to automatically prioritize roles and relationships for selection.

Topic Themes (Harabagiu & Lăcătuşu 2005) used for multi-document summarization merge various arguments corresponding to the same semantic roles for the semantically identical verb phrases (e.g., *arrests* and *placed under arrest*). *Atomic relations* described in Chapter 3 also model an input document collection and are created according to the statistics collected for co-occurrences of named entity pairs linked through actions. GISTexter, atomic relations, and Topic Themes were used for modeling a collection of documents rather than a domain.

In other closely related work, Sudo et al. (2003) use frequent dependency subtrees as measured by *tf*idf* to identify named entities and IE patterns important for a given domain. The goal of their work is to show how the techniques improve IE pattern acquisition. To do this, Sudo et al. constrain the retrieval of relevant documents for a MUC scenario and then use unsupervised learning over descriptions within these documents that match specific types of named entities (e.g., *Arresting Agency, Charge*), thus enabling learning of patterns for specific templates (e.g., the Arrest scenario). In contrast, the goal of our work is to show how similar techniques can be used to learn what information is important for a given domain or event and thus, should be included into the domain template. Our approach allows, for example, learning that an arrest along with other events (e.g., attack) is often part of a terrorist event. We do not assume any prior knowledge about domains. A key difference, then, is in the application of our unsupervised learning approach to instances (e.g., documents representing specific terrorist events) within a domain (e.g., terrorism).

There are also differences in our techniques.We demonstrate that frequent subtrees can be used not only to extract specific named entities for a given scenario but also to learn domain-important relations. These relations link domain actions and named entities as well as general nouns and words belonging to other syntactic categories. We also show that these relations can direct the sentence selection process to generate descriptions of unseen events in the described domains.

Collier (1998) proposed a fully automatic method for creating templates for information extraction. The method relies on Luhn's idea (1957) of locating statistically significant words in a corpus and uses those to locate the sentences in which they occur. Then it extracts Subject-Verb-Object patterns in those sentences to identify the most important interactions in the input data. The system was constructed to create MUC templates for *terrorist attacks*. The idea of using a particular syntactic pattern is similar to the idea applied for (Riloff 1996, Riloff & Schmelzenbach 1998). Our work also relies on corpus statistics, but we utilize arbitrary syntactic patterns and explicitly use multiple domain instances. Keeping domain instances separated, we cross-examine them and estimate the importance of a particular information type in the domain.

4.2 Our Approach to Template Induction

After reading about presidential elections in different countries on different years, a reader has a general picture of this process. Later, when reading about a new presidential election, the reader already has in her mind a set of questions for which she expects answers. This process can be called *domain modeling*. The more instances of a particular domain a person has seen, the better understanding she has about what type of information should be expected in an unseen collection of documents discussing a new instance of this domain.

Thus, we propose to use a set of document collections describing different instances within one domain to learn the general characteristics of this domain. These characteristics can be then used to create a domain template. We test our system on four domains: airplane crashes, earthquakes, presidential elections, terrorist attacks.

4.3 Data Description

4.3.1 Training Data

To create training document collections we used BBC Advanced Search[4] and submitted queries of the type ⟨*domain title* + *country*⟩. For example, ⟨"presidential election" USA⟩.

In addition, we used BBC's Advanced Search date filter to constrain the results to different date periods of interest. For example, we used known dates of elections and allowed a search for articles published up to five days before or after each such date. At the same time for the terrorist attacks or earthquakes domain the time constraints we submitted were the day of the event plus ten days.

Thus, we identify several instances for each of our four domains, obtaining a document collection for *each* instance. E.g., for the earthquake domain we collected documents on the earthquakes in Afghanistan (March 25, 2002), India (January 26, 2001), Iran (December 26, 2003), Japan (October 26, 2004), and Peru (June 23, 2001). Using this procedure we retrieve training document collections for 9 instances of airplane crashes, 5 instances of earthquakes, 13 instances of presidential elections, and 6 instances of terrorist attacks.

4.3.2 Test Data

To test our system, we used document clusters from the Topic Detection and Tracking (TDT) corpus (Fiscus et al. 1999). Each TDT topic has a topic label, such as *Accidents* or *Natural Disasters*.[5] These categories are broader than our domains. Thus, we manually filtered the TDT topics relevant to our four training domains (e.g., Accidents matching Airplane Crashes). In this way, we obtained TDT document clusters for 2 instances of airplane crashes, 3 instances of earthquakes, 6 instances of presidential elections and 3 instances of terrorist attacks. The number of the documents corresponding to the instances varies greatly (from two documents for one of the earthquakes up to 156 documents for one of the terrorist attacks). This variation in the number of documents per topic is typical for the TDT corpus. Many of the current approaches of domain modeling collapse together different

[4] `http://news.bbc.co.uk/shared/bsp/search2/advanced/news_ifs.stm`

[5] In our experiments we analyze TDT topics used in TDT-2 and TDT-4 evaluations.

instances and make the decision on what information is important for a domain based on this generalized corpus (Collier 1998, Barzilay & Lee 2003, Sudo, Sekine & Grishman 2003). We, on the other hand, propose to cross-examine these instances keeping them separated. Our goal is to eliminate dependence on how well the corpus is balanced and to avoid the possibility of greater impact on the domain template of those instances which have more documents.

4.4 Inducing Templates

In this work we build domain templates around verbs which are estimated to be important for the domains. Using verbs as the starting, point we identify semantic dependencies within sentences. In contrast to deep semantic analysis (Fillmore & Baker 2001, Gildea & Jurafsky 2002, Pradhan, Ward, Hacioglu, Martin & Jurafsky 2004, Harabagiu & Lăcătuşu 2005, Palmer et al. 2005) we rely only on corpus statistics. We extract the most frequent syntactic subtrees which connect verbs to the words used in the same subtrees. These subtrees are used to create domain templates.

For each of the four domains described in Section 4.3, we automatically create domain templates using the following algorithm.

Step 1: *Estimate what verbs are important for the domain under investigation.* We initiate our algorithm by calculating the probabilities for all the verbs in the document collection for one domain — e.g., the collection containing all the instances in the domain of airplane crashes. We discard those verbs that are stop words (Salton 1971). To take into consideration the distribution of a verb among different instances of the domain, we normalize this probability by its *VIF* value (verb instance frequency), specifying in how many domain instances this verb appears.

$$Score(vb_i) = \frac{count_{vb_i}}{\sum_{vb_j \in comb\ coll} count_{vb_j}} \times VIF(vb_i) \tag{4.1}$$

$$VIF(vb_i) = \frac{\#\ of\ domain\ instances\ containing\ vb_i}{\#\ of\ all\ domain\ instances} \tag{4.2}$$

These verbs are estimated to be the most important for the combined document collection for all the domain instances. Thus, we build the domain template around these verbs. Here are the top ten verbs for the *terrorist attack* domain:

> *killed*, *told*, *found*, *injured*, *reported*,
> *happened*, *blamed*, *arrested*, *died*, *linked*.

Step 2: *Parse those sentences which contain the top 50 verbs.* After we identify the 50 most important verbs for the domain under analysis, we parse all the sentences in the domain document collection containing these verbs with the Stanford syntactic parser (Klein & Manning 2002).

Step 3: *Identify most frequent subtrees containing the top 50 verbs.* A domain template should contain not only the most important actions for the domain, but also the entities that are linked to these actions or to each other through these actions. The words referring to such entities can potentially be used within the domain template slots. Thus, we analyze those portions of the syntactic trees which contain the verbs themselves plus other words used in the same subtrees as the verbs. To do this we use FREQuent Tree miner.[6] This software is an implementation of the algorithm presented by (Abe, Kawasoe, Asai, Arimura & Arikawa 2002, Zaki 2002), which extracts frequent ordered subtrees from a set of ordered trees. Following (Sudo et al. 2003) we are interested only in the words which are near neighbors of the most frequent verbs. Thus, we look only for those subtrees which contain the verbs themselves and from four to ten tree nodes, where a node is either a syntactic tag or a word with its tag. We analyze not only NPs which correspond to the subject or object of the verb, but other syntactic constituents as well. For example, PPs can potentially link the verb to locations or dates, and we want to include this information into the template. Table 4.1 contains a sample of subtrees for the *terrorist attack* domain mined from the sentences containing the verb *killed*. The first column of Table 4.1 shows how many nodes are in the subtree.

Step 4: *Substitute named entities with their respective tags.* We are interested in analyzing a whole domain, not just an instance of this domain. Thus, we substitute all the named entities with their respective tags, and all the exact numbers with the tag NUMBER. We

[6] `http://chasen.org/~taku/software/freqt/`

nodes	subtree
8	(SBAR(S(VP(VBD killed)(NP(QP(IN at))(NNS people)))))
8	(SBAR(S(VP(VBD killed)(NP(QP(JJS least))(NNS people)))))
5	(VP(ADVP)(VBD killed)(NP(NNS people)))
6	(VP(VBD killed)(NP(ADJP(JJ many))(NNS people)))
5	(VP(VP(VBD killed)(NP(NNS people))))
7	(VP(ADVP(NP))(VBD killed)(NP(CD 34)(NNS people)))
6	(VP(ADVP)(VBD killed)(NP(CD 34)(NNS people)))

Table 4.1: Sample subtrees for the *terrorist attack* domain.

speculate that subtrees similar to those presented in Table 4.1 can be extracted from a document collection representing any instance of a terrorist attack, with the only difference being the exact number of causalities. Later, however, we analyze the domain instances separately to identity information typical for the domain. The procedure of substituting named entities with their respective tags previously proved to be useful for various tasks (Barzilay & Lee 2003, Sudo et al. 2003, Filatova & Prager 2005). To get named entity tags we used BBN's IdentiFinder (Bikel et al. 1999).

Step 5: *Merge together the frequent subtrees.* Finally, we merge together those subtrees which are identical according to the information encoded within them. This is a key step in our algorithm which allows us to bring together subtrees from different instances of the same domain. For example, the information rendered by all the subtrees from the bottom part of Table 4.1 is identical. Thus, these subtrees can be merged into one which contains the longest common pattern:

(VBD killed)(NP(NUMBER)(NNS people))

After this merging procedure we retain for further consideration only those subtrees that are mentioned in each domain instance at least twice. At this step, we make sure that we keep in the template only the information which is generally important for the domain rather than only for a fraction of instances in this domain. We also remove all the syntactic tags as we want to make this pattern as general for the domain as possible. A pattern without syntactic dependencies contains a verb together with a prospective template slot corresponding to this verb:

killed: (NUMBER) (NNS people)

In the above example, the prospective template slots appear after the verb *killed*. In other cases, the domain slots appear in front of the verb. Two examples of such slots, for the *presidential election* and *earthquake* domains, are shown below:

(PERSON) won
(NN earthquake) struck

The above examples show that it is not enough to analyze only named entities, general nouns contain important information as well. We term the structure consisting of a verb together with the associated slots a *slot structure*. Here is a part of the slot structure we get for the verb *killed* after cross-examination of the *terrorist attack* instances:

killed (NUMBER) (NNS people)
(PERSON) killed
(NN suicide) killed

Step 6: *Creating domain templates.* After we get all the frequent subtrees containing the top 50 domain verbs, we merge all the subtrees corresponding to the same verb and create a slot structure for every verb as described in Step 5. The union of such slot structures created for all the important verbs in the domain is called the *domain template*. From the created templates we remove the slots which are used in all the domains. For example,

(PERSON) told.

□

The presented algorithm can be used to create a template for any domain. It does not require pre-defined domain or world knowledge. We learn domain templates from cross-examining document collections describing different instances of the domain of interest.

4.5 Discussion

The above domain template creation procedure is driven by the verbs that are automatically identified as important. Many other systems also rely on verbs while analyzing a semantic structure of domains. For example, many case frames are constructed around verbs (Lehnert et al. 1994, Riloff & Schmelzenbach 1998). Examples of several case frames are presented in Figure 4.2. A recent initiative on adding semantic tags to PennTreeBank (Palmer et al. 2005) also relies on verbs. The updated PennTreeBank is called PropBank and constructed

Roleset kill.01 "cause to die":

Arg0:*killer*
Arg1:*corpse*
Arg2:*instrument*

Figure 4.3: PropBank annotation guidelines for the verb *to kill*.

Template slots created by our system, case frames, and PropBank verbs annotation standard, all capture information about arguments relevant to verbs. The type of this information, however, is different for all three cases.

Case frames list semantic classes related to each verb. Each semantic class has a list of syntactic patterns that can capture nouns corresponding to this semantic class. The list of semantic classes of elements important for a domain is created manually. Thus, moving to a new domain, requires manual work on identifying semantic classes for this domain and creating lists of seed elements for each semantic class.

PropBank annotation standard lists semantic arguments that can be linked to each verb as well as semantic classes that are assigned to each semantic argument. In contrast to case frames, the PropBank semantic classes do not have lexicons that are linked to them. To overcome this issue, it is possible to use WordNet as it was used in GISTexter (Harabagiu

[7] `http://www.cs.rochester.edu/~gildea/Verbs/`

& Maiorano 2002). In this case, however, the system relies on the relations encoded in WordNet and this is not such a good solution if the semantic arguments are verbalized mainly though named entities.

Template slots created as described in Section 4.4 do not have either Case Frame syntactic information or PropBank semantic information. The information encoded in the template slots is more of a lexical nature. For example, the information encoded in the *terrorist attack* domain template slot for the verb *killed* shows that this verb is linked to the noun *people* and to a *NUMBER*. The template slot does not have information that the noun phrase (*NUMBER people*) is a direct object and can be interpreted at *Arg1: corpse*, instead, it gives the exact noun that follows the verb *killed* and is highly related to this verb.

The information encoded in case frames, PropBank, and template slots is of different nature. The ideal situation would be to bring together semantic, syntactic and lexical information. A similar idea lies behind one of the most recent corpus creation initiatives OntoNotes (Hovy, Marcus, Palmer, Pradhan, Ramshaw & Weischedel 2006) that focuses on a domain independent representation of literal meaning that includes predicate structure, word sense, ontology linking, and coreference. Combining semantic, syntactic and lexical information for domains is an interesting research task, however, it is outside of the scope of this thesis.

The use of syntactic information in the domain template induction procedure guarantees that the elements of the selected relations belong to the same syntactic subtree and thus were used within at least the same simple clause.

4.6 Evaluation

The task we deal with is new and there is no well-defined and standardized evaluation procedure for it. Sudo et al. (2003) evaluated how well their IE patterns captured named entities of three pre-defined types. We are interested in evaluating how well we capture the major actions as well as their constituent parts.

There is no set of domain templates which are built according to a unique set of principles against which we could compare our automatically created templates. Thus, we need to create a gold standard. In Section 4.6.1, we describe how the gold standard was created. Then, in Section 4.6.2, we evaluate the quality of the automatically created templates by extracting clauses corresponding to the templates and verifying how many answers from the questions in the gold standard are answered by the extracted clauses.

We considered using PropBank to provide a gold standard, but as described in Section 4.5, PropBank frame slots assign a semantic role to each slot, while our algorithm gives either the type of the named entity that should fill in this slot or puts a particular noun into the slot (e.g., ORGANIZATION, earthquake, people). As mentioned in Section 4.5, an ideal domain template should include semantic information but this problem is outside of the scope of this dissertation.

4.6.1 Stage 1. Information Included into Templates: Interannotator Agreement

To create a gold standard we asked people to create a list of questions which indicate what is important for the domain description. We decided to ask annotators to create lists of questions rather than domain templates because: first, not all of our subjects are familiar with the field of IE and thus, do not necessarily know what an IE template is; second, our goal for this evaluation is to estimate interannotator agreement for capturing the important aspects for the domain and not how well the subjects agree on the template structure.

We asked our subjects to think of their experience of reading newswire articles about various domains.[8] Based on what they remember from this experience, we asked them to come up with a list of questions about a particular domain. We asked them to come up with at most 20 questions covering the information they will be looking for, given an unseen news article about a new event in the domain. We did not give them any input information about the domain but allowed them to use any sources to learn more information about the domain.

[8] We thank Rod Adams, Cosmin-Adrian Bejan, Sasha Blair-Goldensohn, Cyril Cerovic, David Elson, David Evans, Ovidiu Fortu, Agustin Gravano, Lokesh Shresta, John Yundt-Pacheco and Kapil Thadani for the submitted questions.

We had ten subjects, each of which created one list of questions for one of the four domains under analysis. Thus, for the *earthquake* and *terrorist attack* domains we got two lists of questions; for the *airplane crash* and *presidential election* domains we got three lists of questions.

After the questions lists were created we studied the agreement among annotators on what information they consider is important for the domain and thus, should be included in the template. We matched the questions created by different annotators for the same domain. For some of the questions we had to make a judgement call on whether it is a match or not. For example, the following question created by one of the annotators for the *earthquake* domain was:

> *Did the earthquake occur in a well-known area for earthquakes (e.g. along the San Andreas fault), or in an unexpected location?*

We matched this question to the following three questions created by the other annotator:

> *What is the geological localization?*
> *Is it near a fault line?*
> *Is it near volcanoes?*

The choice of the metric that should be used to measure interannotator agreement (e.g., Cohen's Kappa, weighted Kappa with Fleiss-Cohen quadratic weight, polychoric correlation, etc.) depends on the nature of the experiment. For our experiment, we do not ask users to evaluate a pre-defined set of elements. Rather, we ask the annotators to provide a set of elements that are important for a domain. Therefore, we cannot use agreement metrics, like kappa, that assume that users evaluate as relevant or not a set of predefined elements. Hence, we use set-similarity metrics to measure the agreement of the annotators. Such metrics do not require knowledge of the expected or chance agreement. A clean and unbiased metric of set similarity is the Jaccard metric, defined as:

$$Jaccard(domain^d) = \frac{|QS_i^d \cap QS_j^d|}{|QS_i^d \cup QS_j^d|} \tag{4.3}$$

	Jaccard metric		
Domain	subj_1 and subj_2 (and subj_3)	subj_1 and MUC	subj_2 and MUC
Airplane crash	0.54	-	-
Earthquake	0.68	-	-
Presidential Election	0.32	-	-
Terrorist Attack	0.50	0.63	0.59

Table 4.2: Creating gold standard. Jaccard metric values for interannotator agreement.

where QS_i^d and QS_j^d are the sets of questions created by subjects i and j for domain d. For the *airplane crash* and *presidential election* domains we averaged the three pairwise Jaccard metric values.

Table 4.2 shows the values of Jaccard metric for interannotator agreement computed for all four domains.[9]

The scores in Table 4.2 show that for some domains the agreement is quite high (e.g., *earthquake*), while for other domains (e.g., *presidential election*) it is twice as low. This difference in scores can be explained by the complexity of the domains and by the differences in understanding of these domains by different subjects. The scores for the *presidential election* domain are predictably low as in different countries the roles of presidents are very different: in some countries the president is the head of the government with a lot of power, while in other countries the president is merely a ceremonial figure. In some countries the presidents are elected by general voting while in other countries, the presidents are elected by parliaments. These variations in the domain cause the subjects to be interested in different issues of the domain. Another issue that might influence the interannotator agreement is the distribution of the presidential election process in time. For example, one

[9] We could also use other metrics that measure set similarity, such as the cosine similarity. In this case, we could treat each annotator-provided question as a single word and treat the set of questions as a bag-of-words. We decided to use the Jaccard metric because both metrics essentially rely on the ratio of the intersection over the union of the questions, and Jaccard is much easier to interpret than the cosine similarity metric. Furthermore, our initial experiments showed that the two metrics are strongly correlated.

of our subjects was clearly interested in the pre-voting situation, such as debates between the candidates, while another subject was interested only in the outcome of the presidential election.

For the *terrorist attack* domain we also compared the lists of questions we got from our subjects with the *terrorist attack* template created by experts for the MUC competition. In this template we treated every slot as a separate question, excluding the first two slots which captured information about the text from which the template fillers were extracted and not about the domain. The results for this comparison are included in Table 4.2.

Differences in domain complexity were studied by IE researchers. Bagga (Bagga 1997) suggests a classification methodology to predict the syntactic complexity of the domain-related facts. Huttunen et al. (Huttunen, Yangarber & Grishman 2002) analyze how component sub-events of the domain are linked together and discuss the factors which contribute to the domain complexity.

4.6.2 Stage 2. Quality of the Automatically Created Templates

In Section 4.6.1 we showed that not all domains are equal. For some domains it is much easier to reach a consensus about what slots should be present in the domain template than for others. In this section we describe the evaluation of the four automatically created domain templates (Section 4.4).

Automatically created templates consist of slot structures and are not easily readable by human annotators. Thus, instead of direct evaluation of the template quality, we evaluate the sentences extracted according to the created templates and check whether these sentences contain the answers to the questions created by the subjects during the first stage of the evaluation. We extract sentences corresponding to the test instances according to the following procedure:

1. Break all the documents corresponding to a particular test instance (respective TDT topic) into sentences.

2. For every domain template slot check all the sentences in the instance (TDT topic) under analysis. Find the shortest sentence which includes both the verb and other

words extracted for this slot in their respective order. Add this sentence to the list of extracted sentences unless this sentences has been already added to this list.

3. Keep adding sentences to the list of extracted sentences till all the template slots are analyzed or the size of the list exceeds 20 sentences.

The key step in the above algorithm is Step 2. By choosing the shortest sentence corresponding to a particular template slot, we reduce the possibility of adding more information to the output than is necessary to cover this particular slot. However, even in this case the extracted sentence is likely to cover more information than is encoded within the respective template slot. Thus, within each extracted sentence we highlight the part that corresponds to the template slot.

In Step 3 we keep only the first twenty sentences so that the size of the output which potentially contains an answer to the question of interest is not larger than the number of questions provided by each subject. The templates are created from the slot structures extracted for the top 50 verbs. The higher the estimated score of the verb (Eq. 4.1) for the domain the closer to the top of the template the slot structure corresponding to this verb will be. We assume that the important information is more likely to be covered by the slot structures that are placed near the top of the template.

To map the sentences extracted for each document collection (domain instance) onto the question sets created by our annotators at Stage 1 of our evaluation (Section 4.6.1) we used the Amazon Mechanical Turk service.[10] This service offers access to a community of human subjects; furthermore, the service provides tools for distributing small tasks and recruiting human subjects to perform these tasks. For our evaluation study, each Mechanical Turk annotator had to read a set of sentences extracted for a particular domain instance and a set of questions created for this domain. Question sets were created by multiple annotators: for the *Airplane Crash* and *Presidential Election* domains we received three sets of questions from our annotators; while for the *Terrorist Attack* and *Earthquake* domains we received two sets of questions from our annotators. Thus, we submitted multiple cases of the same domain instance sentences to map these sentences to each of the question sets: three sets of

[10] `http://www.mtruk.com`

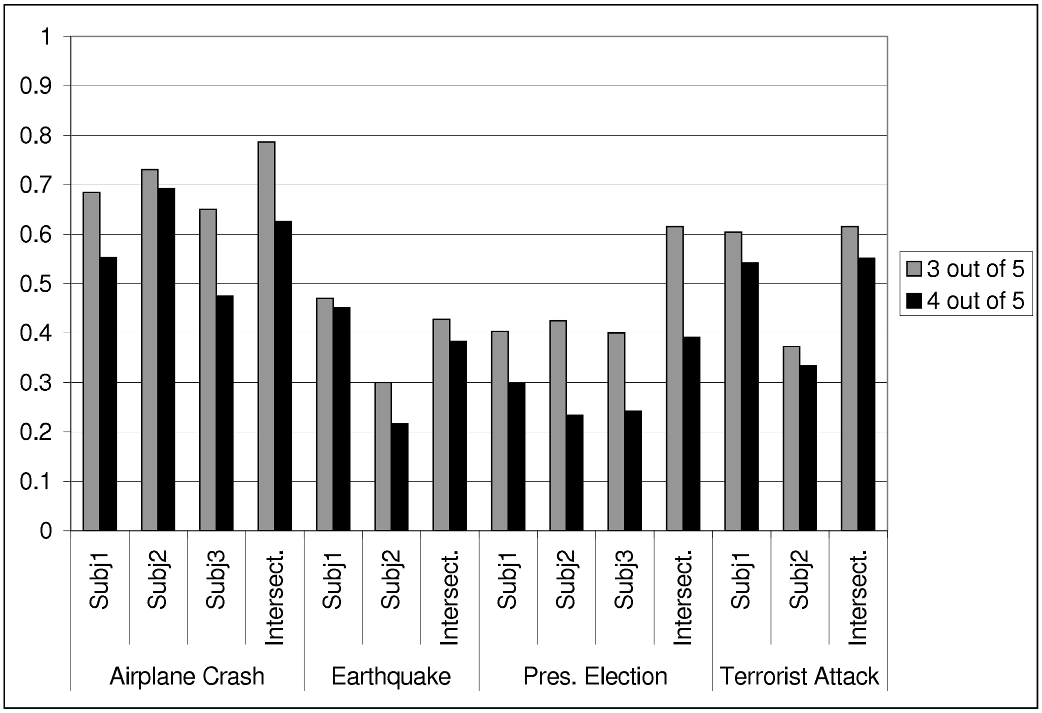


Figure 4.4: Evaluation results.

sentences – questions pairs for each instance of the *Airplane Crash* and *Presidential Election* domains; and two sets of sentences – questions pairs for each instance of the *Airplane Crash* and *Presidential Election* domains. We asked Mechanical Turk annotators to pair each question with the sentences that contain an answer to this question. If none of the extracted sentences provided an answer to a particular question then we asked them to mark such questions with the tag NONE. For each domain instance we provided Mechanical Turk annotators with a brief description of this instance (topic explication from the TDT corpus). For each document collection and list of questions pair, we recruited five annotators. We asked Mechanical Turk annotators to look for answers only within the highlighted portions of domain instance sentences and use the remaining text as context. To take into consideration the nature of the Mechanical Turk annotators and possible interannotator agreement we considered that domain instance sentences contained an answer to a particular question only if the majority of the annotators paired this question with at least one sentence. We computed the performance of our system using two values for majority: three and four.

The evaluation results for the automatically created templates are presented in Figure 4.4. To estimate the performance of our system we compute the average ratio of the questions covered by the outputs created according to the domain templates, i.e., the number of questions that are answered by the produced text. For every domain, we present the ratio of the covered questions separately for each annotator and for the intersection of questions (Section 4.6.1). We present two sets of results: one set is for the case where we considered a particular question answered by the extracted sentences if at least *three out of five* Mechanical Turk annotators paired this question with at least one of the extracted sentences; the other set is for the case where we considered a particular question answered by the extracted sentences if at least *four out of five* Mechanical Turk annotators paired this question with at least one of the extracted sentences.

For the questions common for all the annotators we capture at least 60% of the answers for three out of four domains. After studying the results we noticed that for the *earthquake* domain some questions did not result in a template slot and thus, could not be covered by the extracted clauses. Here are two of such questions:

> *Is it near a fault line?*
> *Is it near volcanoes?*

According to the template creation procedure, which is centered around verbs, the chances that extracted clauses would contain answers to these questions are low. Indeed, only one of the three sentence sets extracted for the three TDT earthquake topics contain an answer to one of these questions.

Poor results for the *presidential election* domain could be predicted from the Jaccard metric value for interannotator agreement (Table 4.2). There is considerable discrepancy in the questions created by human annotators which can be attributed to the great variation in the *presidential election* domain itself. It must be also noted that most of the questions created for the *presidential election* domain were clearly referring to the democratic election procedure, while some of the TDT topics categorized as *Elections* were about either election fraud or about opposition taking over power without the formal resignation of the previous president.

Overall, this evaluation shows that using automatically created domain templates we extract sentences which contain a substantial part of the important information expressed in questions for that domain. For those domains which have small diversity our coverage is significantly higher.

4.7 IE and Automatic Domain Template Induction

In Section 3.6 we gave an overview of the field of IE and pointed our that there exist two approaches to tackle the IE rigid constraint of requiring as input a set of pre-defined data types:

1. extraction of ***all potentially important*** relations or events from the input corpora;
2. unsupervised learning of the relations, events, named entity types that are ***important for a domain***.

In this dissertation we explore both approaches. Atomic relations (Chapter 3) correspond to the first one, while automatically induced templates correspond to the second one.

The data-driven procedure for automatic domain template induction is described in this chapter. We noted that the induced domain templates do not contain rich semantic and syntactic information about the domain relations as manually created MUC templates (Marsh & Perzanowski 1997), case frames (Riloff & Schmelzenbach 1998), PALKA's frame-phrasal pattern structures (Kim & Moldovan 1995), or PropBank annotation schemas (Palmer et al. 2005). At the same time, the automatically induced templates contain information about the exact words that can be used to express semantic arguments linked to the verbs identified as important for the domain. Thus, the domain templates that we automatically induce according to the procedure described in this chapter are not as accurate as semantically full-fledged domain templates manually created by domain experts. Rather, our templates contain information about what words or types of named entities are closely related to the verbs important for a particular domain.

The domain slots (either stand-alone slots or groups of slots corresponding to the same verb) can be used as initial patterns for capturing relations. There are a lot on-going efforts for analyzing semantic arguments of verbs and grouping together noun phrases corresponding to the identical semantic arguments (Harabagiu & Lăcătuşu 2005, Qiu, Kan & Chua 2006). If the elements linked to the verbs can be assigned appropriate semantic classes and the induced domain templates can be enhanced with semantic tags for the arguments, then the induced domain templates have a potential to be as well-defined as manually created templates. In addition, automatically induced templates would contain one of the patterns (a seed pattern) which can be used for filling in the templates. This seed pattern can be used to start a bootstrapping mechanism similar to the one use in Snowball (Agichtein & Gravano 2000) in order to start the cycle of getting tuples corresponding to the input relation and patterns capable of capturing these tuples.

4.8 Conclusions

In this chapter, we presented a robust method for data-driven discovery of the important fact-types for a given domain. In contrast to supervised methods, the fact-types are not prespecified. The resulting slot structures can subsequently be used to guide the generation of responses to questions about new instances of the same domain. Our approach features the use of corpus statistics derived from both lexical and syntactic analysis across documents. A comparison of our system output for four domains of interest shows that our approach can reliably predict the information that humans have indicated are of interest. Our method is flexible: analyzing document collections from different time periods or locations, we can learn domain descriptions that are tailored to those time periods and locations.

In this chapter we showed how to induce automatically commonalities across different instances of a particular domain. In Chapter 6, we show how commonalities can be established across different domains. We describe the design of the procedure that we developed for identifying commonalities across different subdomains (in our example, across activities that could be used for descriptions of people belonging to different occupations). We show how to use random walk theory to identify biographical information that corresponds to three levels of activity. Our unsupervised learning technique identifies automatically three levels of activities: general biographic, occupation-related and person-specific. We show

how the identified domain commonalities can be used to create domain hierarchies.

Chapter 5

Columbia Event QA System and its Evaluation

> There are two sides to every question.
>
> Protagoras (485 BC - 421 BC)

Many current QA systems focus on answering various types of questions requiring a text as an answer (open-ended questions) rather than a short text snippet (factoid questions). A number of approaches have been proposed recently to answer open-ended questions, such as biography questions (Blair-Goldensohn, Evans, Hatzivassiloglou, McKeown, Nenkova, Passonneau, Schiffman, Schlaikjer, Siddharthan & Siegelman 2004, Zhou et al. 2004), definition questions (Blair-Goldensohn, McKeown & Schlaikjer 2004), and opinion questions (Stoyanov, Cardie & Wiebe 2005). In this thesis, we suggest and evaluate a two-pronged approach for answering event-related open-ended questions.

QA systems answering open-ended questions of any type have a goal of identifying portions of text within document collection that should be included in the answer. The solutions for identifying such portions of text differ for different question types. For example, in definition or biography questions, the information about the person, object, or notion in question is highly likely to be drawn from the sentences containing the question term or a reference to this term, or from sentences that are in close proximity to the question term (person, object or notion). In event-related questions, however, the situation is quite different. On the one hand, descriptions of events in question can be long and complex, and thus, choosing a term central for an answer is not trivial; on the other hand, many relevant sentences that should be included in the answer do not contain any of the question terms. For example, in questions of the type "DESCRIBE THE PRESIDENTIAL ELECTION IN COUNTRY *X* IN YEAR *NNNN*" there is no explicit reference to the candidates' names, campaigning, vote distribution, though it is likely that the information about these aspects should be included in the answer. In another example "LIST FACTS ABOUT EVENT [*The shut down of the Cernavoda nuclear power plant*]," all the sentences about the level of water in the Danube river at Cernavoda village would be relevant, even though these sentences do not mention directly the closure event:

- *The Danube at Cernavoda village, where the reactor is located, fell to a depth of less than three meters (10 feet) on Saturday, down from its usual level of almost seven meters (23 feet).*
- *"If it rains in Western and Central Europe, the increased water takes 25 days to reach Cernavoda," he told national radio, adding that the hot summer had led to a seven-percent increase in energy demands.*

As we analyzed event-related questions we decided to break these questions into two large groups:

1. Questions describing events that are either unique or have a complex verbalization and thus, cannot be generalized and assigned to a particular domain. For example, "LIST FACTS ABOUT EVENT [*The shut down of the Cernavoda nuclear power plan*]".
2. Questions requesting information about an event that can be generalized and assigned to a particular domain. For example, the event in questions of the type "DESCRIBE THE PRESIDENTIAL ELECTION IN COUNTRY *X* IN YEAR *NNNN*" can be assigned to a *presidential election* domain.

Thus, we developed a two-pronged approach for answering event-related questions. For the first type of event-related questions, we rely on a shallow semantic network constructed of atomic relations extracted from the input document collection. For the second type of event-related questions, we use a set of domain-specific relations to guide the process of answer generation. The major goal of both approaches is to find the sentences that should be included in the answer and rank those sentences by the order of importance of the information they cover.

We tested our event QA module as part of the DARPA GALE program.[1] GALE systems deal with several question types that can be classified as event questions:

- Template 1: List facts about event [*event*].
- Template 8: Describe the prosecution of [*person*] for [*crime*].

For Template 1 questions, we used an answer generation approach based on a shallow semantic network. For Template 8 questions, we used the module that selects sentences containing relations identified as important for the prosecution domain. There are two ways to identify these relations: use ACE *justice* events identified by an IE system (Schiffman, McKeown, Grishman & Allan 2007), or use a domain template created according to the procedure described in Chapter 4.

5.1 Related Work

Answering questions formulated in natural language is a necessary functionality to pass the famous *Turing Test* (Turing 1950). The first attempts to tackle the question answering (QA) problem date to the middle of the 20th century (Simmons 1965). QA systems vary in the range of questions they can answer and the corpora they analyze to extract an answer. Some of the first QA systems were natural language interfaces for well-structured knowledge repositories (i.e., data bases) in particular domains (Androutsopoulos, Ritchie & Thanisch 1995), including: *Baseball* (Green, Wolf & amd Kenneth Laughery 1963), *SAD SAM* (Lindsay 1963), *LUNAR* (Woods 1978), *PHILQA1* (Bronnenberg, Bunt, Landsbergen, Scha, Schoenmakers & Utteren 1979). For example, *LUNAR* (Woods 1978) was created

[1] An overview of the GALE program is available at: `http://projects.ldc.upenn.edu/gale/` and at `http://www.darpa.mil/ipto/programs/gale/index.htm`.

to answer questions about chemical data on lunar material compiled during the Apollo missions; *Baseball* (Green et al. 1963) answered questions about the outcomes of baseball games for which information was stored using a predefined format.

Among the early text-based QA systems were *The Automatic Language Analyzer (ALA)* (Thorne 1962), *Protosynthex* (Simmons, Klein & McConlogue 1964), TEXT (McKeown 1985), and a question answering system developed by Wendy Lehnert (Lehnert 1978). *ALA* (Thorne 1962) was designed to handle questions based on an astronomy book, *Protosynthex* (Simmons et al. 1964) was the first attempt to answer questions from an encyclopedia, McKeown's TEXT (1985) used schemas to encode rhetorical knowledge needed to answer different kinds of open-ended questions (e.g., definitions), and the system described in (Lehnert 1978) was designed to analyze the semantics of text and the questions inquired for inferences from semantic structure of the input text.

Although the first QA systems used a wide variety of approaches, there was a number of processing requirements used within most systems. Strikingly, the most successful current QA systems have the same requirements:

- strong organization of data;
- syntactic processing of questions and corpora;
- semantic mapping of questions to potential answers.

Many current QA systems use various structured and semi-structured data repositories for answer detection and validation. These data repositories are usually either publicly available on-line encyclopedias, such as CIA World Factbook, IMDB, Biography.com, Wikipedia.com, etc. (Clarke, Cormack, Lynam, Li & McLearn 2001, Echihabi & Marcu 2003, Lita, Hunt & Nyberg 2004, Katz, Marton, Borchardt, Brownell, Felshin, Loreto, Louis-Rosenberg, Lu, Mora, Stiller, Uzuner & Wilcox 2005, Katz, Borchardt & Felshin 2005), or repositories created for specific question types (Fleischman, Echihabi & Hovy 2003). Mapping of semantic structures of questions and possible answers is also present in most modern QA systems. In some QA systems this mapping procedure involves deep semantic understanding of text (Moldovan, Harabagiu, Gîrju, Morărescu, Lăcătuşu, Novischi, Badulescu & Bolohan 2002, Harabagiu, Moldovan, Clark, Bowden, Hickl & Wang 2005);

other QA systems use shallow semantic information by searching for an answer with the of patterns corresponding to the input question type (Soubbotin & Soubbotin 2001, Ravichandran & Hovy 2002, Jijkoun, Mur & de Rijke 2004); many QA systems combine the advantages of both deep and shallow semantic approaches (Peng, Weischedel, Licuanan & Xu 2005).

5.1.1 Factoid Questions

The first formal evaluations of QA systems were performed within the TREC competitions. The systems participating in TRECs typically were a mixture of Information Retrieval (IR) and Information Extraction (IE). TREC QA evaluations, which were launched by NIST in 1999, set the standard for both the types of questions targeted by the systems and for the evaluation metrics used to judge the quality of the QA systems (Voorhees & Harman 1999, Voorhees & Harman 2000, Voorhees 2001, Voorhees 2002, Voorhees 2003*b*, Voorhees 2004, Voorhees 2005).

Systems participating in TREC deal with open domain factoid short-answer questions (i.e., similar to the ones from *Trivial Pursuit* or *Who wants to be a millionaire* games). The latest TREC systems are also evaluated on definition and list questions. One of the major characteristics of short-answer QA systems is that the type of the answer is clearly identified by the question and the length of the answer is usually a word or a phrase. The first step in the QA process is document retrieval, and though this step is crucial, it has been shown that this step does not provide enough information to produce an answer to the input question (Kwok, Grunfeld, Dinstl & Chan 2000, Radev, Prager & Samn 2000). Thus, in addition to document retrieval, a wide variety of NLP techniques is used by factoid QA systems.

One of the pioneering systems dealing with factoid QA was MURAX (Kupiec 1993). It used an on-line encyclopedia to answer a closed class of questions expecting either a named entity or a number to be an answer. Another QA system, developed by Srihari and Li (Srihari & Li 2000), heavily relied on NE tagging, as many factoid questions were asking for a named entity (e.g., who questions, where questions, etc.). Cardie *et al* (Cardie, Ng, Pierce & Buckley 2000) used a combination of passage vector representation models together with query-dependent summarization, and shallow syntactic and semantic sentence analysis for pointing out text snippets containing answers. There are QA systems that use question reformulation and abductive proofs to map the question logical forms onto the logical forms which potentially contain answers (Harabagiu, Moldovan, Paşca, Mihalcea, Surdeanu, Bunescu, Gîrju, Rus & Morărescu 2001). A graph-based approach for mapping question logical forms onto text snippets potentially containing answer logical forms was suggested in (Mollá & van Zaanen 2005). Another approach to factoid QA is to rely on redundancy as it was shown that the more occurrences of a correct answer exists in the corpus, the better the chances of a QA system to extract this correct answer (Light, Mann, Riloff & Breck 2001). This estimation was used in the systems that employed data-intensive techniques and used the web as their corpus (Brill, Dumais & Banko 2002, Lin 2007).

An approach that recently has become very popular among factoid QA systems relies on sets of patterns that are either manually constructed or automatically learned for each question type (Soubbotin & Soubbotin 2001, Ravichandran & Hovy 2002, Echihabi, Hermjakob, Hovy, Marcu, Melz & Ravichandran 2003, Chu-Carroll, Prager, Welty, Czuba & Ferrucci 2003, Greenwood & Saggion 2004, Jijkoun et al. 2004).

A number of recent NLP systems deal with pre-computing large data repositories which can be potentially used for QA (Fleischman et al. 2003, Etzioni et al. 2004, Cafarella, Downey, Soderland & Etzioni 2005, Paşca, Lin, Bigham, Lifchits & Jain 2006). The importance of having such data repositories was proven by the analysis presented in (Agichtein, Cucerzan & Brill 2005). This analysis showed that answers to many factoid questions can be stored in a small set of structured repositories corresponding to various relations.

5.1.2 Questions Requiring Long Answers

Recently, many QA systems moved from factoid questions to questions requiring long detailed answers. Systems participating in the latest TREC competitions created answers to questions of the type *"Other"* (Voorhees 2005). In the recent TREC competitions all the questions are centered around *targets*, where a target is a name of a person, a notion, an event, etc. A set of factoid and list questions is formulated for each target. Each target also has a question of the type *"Other"* that requires as an answer a set of text snippets containing information important for the target but not covered by the factoid and list questions for this target. The combination of the text snippets corresponding to the *"Other"* question can be considered as a part of an answer to a definition question "What is *target*?"

The 2003 TREC competition also introduced definition questions (Voorhees 2003*b*) and the 2004 Document Understanding Conference (DUC) introduced topic-centered summarization; in particular, DUC 2004 asked the participating systems to tailor their systems to answer "Who is X?" questions (Blair-Goldensohn, Evans, Hatzivassiloglou, McKeown, Nenkova, Passonneau, Schiffman, Schlaikjer, Siddharthan & Siegelman 2004, Zhou et al. 2004). The QA systems targeting long answers borrow many techniques developed by factoid QA systems, like pre-computed patterns (Blair-Goldensohn, McKeown & Schlaikjer 2003, Hang, Kan & Chua 2005) or training towards extracting pre-defined types of information (Zhou et al. 2004).

As noted, QA systems answering open-ended questions borrow many techniques from summarization, especially focused summarization (like DUC 2004 biography generation). In addition, a number of novel summarization techniques were introduced to make the summaries more focused. For example, a summarization system based on Topic Themes (Harabagiu & Lăcătuşu 2005) analyzes semantically identical verb phrases and merges various arguments corresponding to the same semantic roles of these verb phrases. Techniques based on semantic parsing are becoming more and more popular due to the initiative of incorporating semantic tags into the PennTree Bank (Kingsbury & Palmer 2002, Palmer et al. 2005). GISTexter (Harabagiu & Maiorano 2002) creates topic templates for the input document collection relying on the relations encoded in WordNet. The system created by BBN for the TREC definition questions (Xu, Licuanan & Weischedel 2003) introduces a notion of *kernel facts* that combines patterns and pre-defined information types but also uses the relations and events as marked for the ACE evaluation (Doddington et al. 2004).

It has become more and more important to extract information not only about certain objects but also about the relations among these objects (Harabagiu 2004, Narayanan & Harabagiu 2004). Similar goals are formulated by Guha *et al* (2003) who describe semantic search over the semantic web where "the semantic web is not a web of documents, but a web of relations between resources denoting real world objects." They show an advantage of having real world objects linked to each other. For example, asking about ⟨**Yo-Yo Ma**⟩ we get a network of objects. This network encodes information that Yo-Yo Ma is a ⟨**musician**⟩, he is connected to ⟨**Paris, France**⟩ by the link ⟨**birth place**⟩, to ⟨**10/07/55**⟩ by the link ⟨**birth date**⟩, etc.

Research on the Semantic Web relies on tagged data, existing Knowledge Bases (KB) and ontologies. Most of the successful QA systems rely on pre-computed patterns, detailed knowledge bases and ontologies as well. The assumption of having all the data tagged and stored in various KBs and ontologies is a very strong one. Thus, the necessity of having unsupervised data-driven learning approaches for identifying relations important for the questions that should be included in the answer is crucial. In this thesis, we propose, implement and evaluate several unsupervised relation learning approaches and demonstrate how they can be used to improve QA systems.

5.1.3 The Specifics of Event-Related Questions

The first formally evaluated systems answering event-related questions were the QA systems that participated in the TREC competitions. Event-related questions answered by such systems were factoid questions asking about where and when something happened, who did something, etc. For example,

- Where did Dylan Thomas die? (TREC-8, question 151)
- When did the Hindenburg crash? (TREC-2001, question 1010)
- Who killed John F. Kennedy? (TREC-2001, question 1274)

Factoid QA systems used a variety of techniques to answer such event questions (Section 5.1.1). Systems developed specifically for answering factoid event questions study links among locations (*where*), times (*when*), and event participants and objects (*who* and *what*). One of such systems is QUALIFIER (Yang, Chua, Wang & Koh 2003). It uses link analysis and external knowledge (e.g., Web, WordNet) to answer factoid event-related questions.

Recently, NIST started a new evaluation effort for IE event identification, Automatic Content Extraction (ACE) (Doddington et al. 2004). For the ACE task, the participating

Life event subtype	Arguments
Be Born	Person, Time, Place
Die	Agent, Victim, Instrument, Time, Place
Injury	Agent, Victim, Instrument, Time, Place
Marry	Person, Time, Place
Divorce	Person, Time, Place

Table 5.1: ACE *Life* event subtypes.

systems are supposed to identify several pre-defined semantic types of events (*life*, *justice*, *transaction*, *conflict*, etc.) together with the constituent parts corresponding to these events (*agent*, *object*, *source*, *target*, *time*, *location*, *other*). For example, Table 5.1 lists *life* events together with the arguments which should be extracted for these events. Many of the ACE systems use semantic parsers created according to the PropBank (Kingsbury & Palmer 2002) semantic annotation. PropBank is a recent initiative which adds semantic tags to the PennTreeBank corpus. For example, the *kernel facts* used by the BBN system for the TREC definition questions (Xu et al. 2003) combine patterns and pre-defined information types, but also use the relations and events as marked for the ACE evaluation (Doddington et al. 2004). Schiffman *et al* (2007) used ACE *justice* event subtypes to select sentences that can be used within an answer to open-ended event-related questions requesting information about prosecutions of people for various crimes.

5.2 Columbia QA System for Answering Open-Ended Event-Related Questions

One of the contributions of this dissertation is an implementation of a system for answering event-related open-ended questions using the techniques we explored in the earlier chapters. Within this pipeline we implement novel components for query input, integrated document retrieval and analysis, and a two-pronged procedure for answer sentence selection: using an event-focused summarization system, and using automatically induced domain templates. We then analyze the performance of our system and repurpose the document retrieval and event-focused summarization modules to better suit the QA task. We develop and evaluate three different versions of our system. In the rest of this chapter, we first present a high-level description of the system architecture, and the evaluation framework (test questions and evaluation metrics) within which our system was tested. Then, we describe the first version of our system tracing an actual system run. Then, we provide details about the error sources and the methods used to target the discovered error sources. We further describe

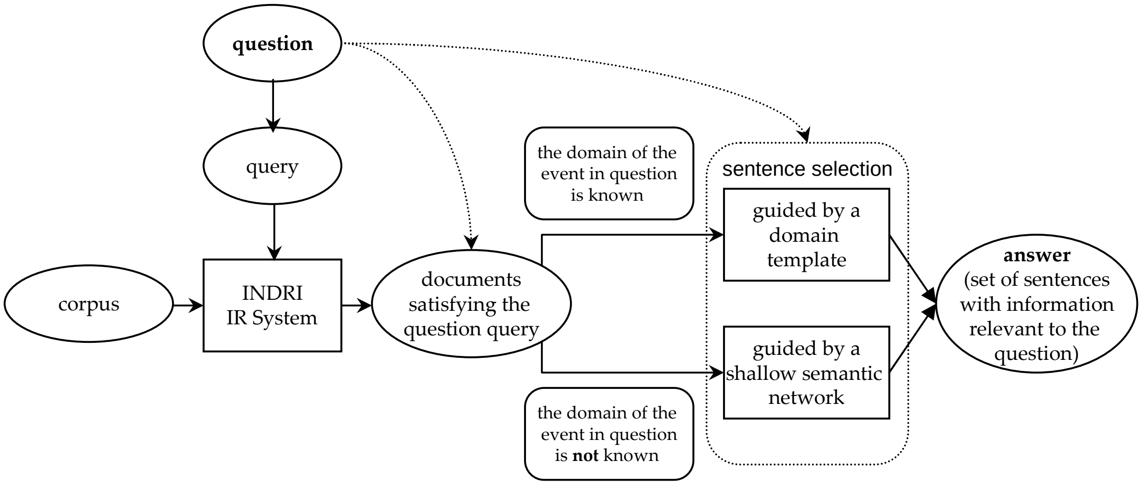


Figure 5.1: Pipeline for Columbia event-related open-ended QA system.

5.2.1 Pipeline Overview

Figure 5.1 illustrates a module-level view of our event-related open-ended QA system pipeline. Our processing proceeds in the following stages:

Query Input and Document Retrieval The goal of the information retrieval component of a QA system is to locate relevant documents that the answer generator can use to construct an answer. For our experiments we use document collections returned by the Indri search engine (Strohman, Metzler, Turtle & Croft 2005). Indri provides a powerful query language that is used to combine numerous aspects of the query and, if possible, utilizes the information related to the event in question. For example, for the questions of the type "DESCRIBE THE PROSECUTION OF [*person X*] FOR [*crime Y*]", the queries submitted to the Indri search engine are enriched with a set of prosecution-related keywords such as: *prosecution*, *defense*, *trial*, *sentence*, *crime*, *guilty*, or *accuse*, all of which were determined on training data to occur in descriptions of prosecutions. The situation with questions of the type "LIST FACTS ABOUT EVENT [*event*]" where it is not possible to predict the domain of the event in question, is different. It is not possible to come up with a set of keywords that are relevant and important for any event. Thus, no query expansion based on a pre-defined language model can be performed for such questions, rather, the corresponding queries submitted to the Indri search engine contain combinations of the terms used in the description of the *event* in question. Each of the terms gets a weight assigned to it.

In addition to query expansion, another important issue that we had to address with respect to the IR stage is the estimation of the number of documents to be returned by Indri. Indri represents queries in probabilistic framework; thus, it is possible that an irrelevant document can be assigned a non-zero score and consequently be considered as relevant to the input query. The scoring is performed on a relative, rather than absolute scale and thus, there exists no common threshold that can be used to cut out irrelevant documents for all questions. However, the returned documents are ranked according to the degree of relevance to the query; the most relevant documents are at the very top of the returned documents list, and the least relevant documents are at the bottom of the list. Thus, it is crucial to correctly estimate how many documents should be requested from Indri. We use different estimation procedures in the three versions of our QA system.

Sentence Selection The sentence selection algorithm depends on the event in question. As noted, we divide all the events in two groups:

1. Events that are either unique or have a complex verbalization and thus, cannot be generalized and assigned to a particular domain.

2. Events that can be generalized and assigned to a particular domain. For example, prosecutions from the questions corresponding to the type "DESCRIBE THE PROSECUTION OF [*person X*] FOR [*crime Y*]".

Thus, we developed a two-pronged approach for answering event-related questions. For the first type of event-related questions, we rely on a shallow semantic network constructed of atomic relations extracted from the input document collection. For the second type of event-related questions, we use a set of domain-specific relations to guide the process of answer generation.

For both sentence selection algorithms the main goal is to include in the answer all the sentences that contain information relevant to the question. In this work we do not try to solve the problem of creating a cohesive well-structured text, rather, we list the selected sentences in the answer according to the importance of the information they contain, with the sentences containing the most important information being in the very beginning of the answer.

5.2.2 Event-Related Questions within GALE

We evaluated our event-related open-ended QA system within the DARPA GALE[2] evaluation program. The systems participating in this program create answers to a set of pre-defined templates. Several of these templates correspond to event-related questions, for example,

- Template 1: "List facts about event [*event*]"
- Template 8: "Describe the prosecution of [*person X*] for [*crime Y*]"

The GALE relevance guidelines,[3] describe what information should be included in answers to the questions corresponding to the GALE pre-defined templates. Figure 5.2 contains the official guidelines for Template 1 questions and Figure 5.3 contains the official guidelines for Template 8 questions.

The guidelines presented in Figure 5.2 and Figure 5.3 are quite generic and are not possible to encode *as is* within an answer generator algorithm. For example, for Template 1, the guidelines ask for information about subevents without giving an exact definition of an event and thus, a subevent. In Chapter 2 we showed that however intuitive the definition of an event is, events identified by various systems vary greatly in both structure

[2] Global Autonomous Language Exploitation

[3] BAE Systems Advanced Information Technologies, "Relevance Guidelines for Distillation Evaluation for GALE: Global Autonomous Language Exploitation," Version 2.2, January 25, 2007.

Relevant information focuses on the event and the persons, places, and activities directly associated with the event. Relevant information also includes subevents, causes, goals, and precursor or preparation events.

For persons associated with the event: relevant information includes their role and the reason for their involvement in the event, as well as their involvement in promoting, funding, or planning for the event. Information about a person must be directly related to the event if it is to be relevant. Involvement in the event may be confirmed or suspected.

Direct reactions, direct consequences, and the significance of the event are also relevant. For locations associated with events: only information that directly pertains to the even is relevant.

Activity date: *indicates date of the event*
Location: *indicates the location of the event*
Categories: *time, location, cause/intention/planning, participant, subevent/execution /manner, consequence/reaction/significance*

Figure 5.2: Official guidelines for Template 1 questions.

and the amount of information they cover. Also, it is easy to see that for neither of the two templates, the use of a few search terms that are explicitly mentioned in the question is sufficient to locate a comprehensive answer. It must be noted, though, that the description of Template 8 can be used to deduce clues about the terms that can be used for query expansion and the answer selection process. For example, one can use ACE *justice* event subtypes to locate text snippets that should be included in an answer for a Template 8 question. The ACE *justice* event subtypes are: arrest, sentence, indict, extradite, charge, execute, release, jail, try, acquit, parole, pardon, hold hearing, fine, sue, convict, appeal. Obviously, most of the ACE *justice* event subtypes can be used to answer GALE Template 8 questions. Like, the ACE *life* event subtypes presented in Table 5.1, ACE *justice* event subtypes have lists of arguments associated with each event subtype.

In the complete version of our QA system, an answer produced by a respective sentence selection procedure is further processed: sentences containing redundant information are grouped together. This issue, however, is outside of the scope of this dissertation.

In the rest of this chapter, we describe our two-pronged approach for answering open-ended event-related questions. To answer Template 1 questions (prong 1), we use a sentence selection algorithm based on the event-focused summarization methodology described in Chapter 3. To answer Template 8 questions (prong 2), we use a sentence selection algorithm

Relevant information includes descriptions of the person's alleged involvement in the crime. Include discussions of the prosecution in describing the activities, motivations, and involvement in the crime. Include descriptions of the person only insofar as they bear directly on the trial, such as the degree to which the person matched the suspect profile. Also include information regarding the defense of the person, and how the defense responded to the allegations of the prosecution.

Include information that describes the sentencing of the person, and related crimes committed (or allegedly committed) by the individual, including similar cases involving the person even when the person was found to be not guilty.

Include reported information believed to be relevant to the case, but deemed inadmissible in a court of law.

Include information such as the country of the trial, the length of the trial, and how the prosecution and defense teams were chosen.

The defendant's arrest for the specific crime is relevant. Also relevant are statements the defendant has made about the specific crime. Reactions of the individuals involved in the trial (jury members, defendants, victims, legal counsel, judges) are relevant. Also relevant are official reactions (such as an official release from a government), general public reactions (rejoicing, outcries, rioting), but statements by other individuals are not relevant.

Activity dates: *Time frame for the prosecution*
Location: *Location for the prosecution*
Categories: *person involved in the crime, crime, victim, defense, prosecution, trial, plea, sentence, reaction, legal process, pretrial process, motivation, defendant statement, related-crime, country-of-trial*

Figure 5.3: Official guidelines for Template 8 questions.

guided by the set of relations pre-defined as related to the *prosecution* event. These relations can be identified within an automatically induced domain template for the *prosecution* domain, or labeled according to the ACE *justice* event subtypes as described in (Schiffman et al. 2007, Hakkani-Tür, Tur & Levit 2007).

5.2.3 Evaluation Set-Up

We evaluate the quality of the created answers for the questions that were used for the first GALE Go/No-Go formal distillation evaluation held in July 2006: seven questions for Template 1 and five questions for Template 8. The answers submitted by our system and two other participating systems were evaluated manually by BAE Systems[4] annotators. Annotators who evaluated the answers of the three systems participating in the GALE

[4] `http://www.alphatech.com/primary/index.htm`

competition used a variety of metrics to analyze the quality of the answers. In our discussion, we use three of those metrics, namely recall, precision and F-measure as these metrics are frequently used to analyze the quality of NLP systems results.

To be able to evaluate the development of our system after introducing changes, we decided to replicate the scores produced by human annotators automatically. We collected all the answer snippets produced by all the three participating systems (including our own system) that were judged as relevant, and the answer snippets produced by human annotators. Combining answer snippets produced by all the systems and human annotators we created answer models for the test questions. To evaluate our answers automatically, we computed cosine similarity between the answer snippets produced by our system and the answer model snippets. If for a sentence selected by our system for the answer there existed a snippet in the respective answer model for which the cosine similarity of the answer sentence and model snippet was higher than 0.7, then we considered that answer sentence relevant to the question, the rest of the sentences were marked as irrelevant. We computed our precision as the ratio of the number of sentences relevant to the question to the number of all the sentences produced by our system for this question (Equation 5.1).

$$\text{Precision}_{questionID} = \frac{\text{Number of relevant answer snippets}(questionID)}{\text{Number of all answer snippets}(questionID)} \tag{5.1}$$

To automatically compute recall value we used an equation similar to the one we used for automatic precision computation. For recall computation, the numerator of the ratio is equal to the number of model answer snippets covered by the automatically produced answer; and the denominator is equal to the overall number of model answer snippets (Equation 5.2).

$$\text{Recall}_{questionID} = \frac{\text{Number of model answer snippets covered by the answer}(questionID)}{\text{Number of all model answer snippets}(questionID)} \tag{5.2}$$

According to the annotation guidelines, the F-measure for the manual annotation was computed with $\beta = 1$. We computed F-measure values using the automatically computed precision and recall values (Equation 5.3).

$$\begin{aligned}\text{F-measure}_{questionID} &= \frac{(1+\beta^2) * \text{Recall}_{questionID} * \text{Precision}_{questionID}}{\text{Recall}_{questionID} + \beta * \text{Precision}_{questionID}} = \\ &= \frac{2 * \text{Recall}_{questionID} * \text{Precision}_{questionID}}{\text{Recall}_{questionID} + \text{Precision}_{questionID}} \qquad (5.3)\end{aligned}$$

In the next section, we show that though the automatically computed precision, recall and F-measure are not equal to their manually computed counterparts they tracks these manually computed counterparts rather well. Thus, can be used to track the changes in different system versions.

5.3 Answering *Template 1* Questions (Prong 1)

In this section, we describe three versions of our event-related open-ended QA system where the answer sentence selection procedure is guided by a shallow semantic network. In Chapter 3, we described the summarization system that we designed and implemented for summarizing event-focused document collections. According to the evaluation results, the performance of this system was high, especially for the document collections that were centered around one event with a well-defined set of constituent parts (e.g., participants, locations, times). Thus, we decided to use this system for creating answers to event-related questions.

We use the Indri search engine to provide us with a set of documents from which an answer to a question can be extracted. As noted, for Template 1 questions,[5] it is not possible to come up with a set of keywords relevant and important for any event. Thus, no query expansion based on a pre-defined language model can be performed for such questions: queries submitted to the Indri search engine contain combinations of the terms used in the description of the *event* in the input question. Each of the terms gets a weight assigned to it. Thus, the Indri search engine performance document retrieval based on a minimal description of the event in question.

To select the optimal set of sentences for a particular question given a document collection, we need to induce the structure of the event in question. Our assumption is, that

[5] Template 1: LIST FACTS ABOUT EVENT [*event*].

an accurate event structure can be used to guide the answer sentence selection process. We use the shallow semantic network extracted from the retrieved document collection as an approximation of the event structure.

The rest of the section contains the descriptions of the three versions of our QA system. Version 1 was created for the Year 1 GALE Go/No-Go evaluation held in July 2006. After we analyzed the Go/No-Go evaluation results, we found two ways to improve the performance of our system. We implemented these extensions consecutively in Version 2 and Version 3 of our event-related open-ended QA system.

5.3.1 System Version 1

5.3.1.1 System Version 1: Information Retrieval Stage

In the first version of our QA system, the number of the documents is estimated according to the number of documents containing named entities mentioned in the *event* in question. If no named entity is mentioned in the *event* in question we ask for ten documents. The parameters for estimation procedure were adjusted empirically.[6] If the number of the requested documents is low (lower than ten), we ask for ten documents to be returned. If the number of requested documents is high (higher than twenty) we ask for twenty documents to be returned. In the first version of our QA system, we treat all the documents returned by the Indri search engine uniformly.

As noted, Indri returns a ranked list of documents with assigned relevance scores. The relevance of documents is computed according to the Indri probabilistic framework and the weights assigned to the terms used in the description of the *event* in question. In many cases, weighting some of the question terms higher than others leads towards a more robust and flexible search. There are questions, however, for which Indri's use of a probabilistic model for assigning weights to the *event* terms causes deterioration of the precision of IR results. For example, for the question

> LIST FACTS ABOUT EVENT [*The shut down of the Cernavoda nuclear power plant*]"

[6] We estimated the search parameters, given twelve training queries.

our estimation procedure requested fourteen documents to be returned. The query consisted of the minimal description of the event in question, namely the lexemes used for the description of the event in the question text. Thus, for the above question, the Indri query consists of various combinations of the lexemes *shut*, *down*, *cernavoda*, *nuclear*, *power*, and *plant*. Out of the fourteen documents returned by Indri, only the top three documents were on topic. The rest of the documents did not contain any relevant information: documents four through eight were on the history of Holocaust in Romania; the ninth document was on Romanian consulate in Hong Kong; the tenth and twelfth documents were on ethnic Hungarians in Romania; the eleventh and thirteenth documents were about Denmark supporting Romania's plea for joining European Union the fourteens document was on human rights violation among gypsies in Romania. However, the corpus that was used to retrieve documents contained overall eleven documents on the subject which means that eight documents were pushed off the IR output list by irrelevant documents.

5.3.1.2 System Version 1: Sentence Selection Stage

To select answer sentences from the Indri IR output we use the shallow semantic network that we construct using atomic relations. The shallow semantic network is constructed following the procedure described in Chapter 3.

1. We identify top ten high frequency nouns in the retrieved documents;
2. We extract all pairs of named entities and high frequency nouns from the in retrieved documents (both pair elements should occur within one sentence);
3. We identify all verbs and action nouns that occur between the elements in the pairs of named entities and high frequency nouns (triples consisting of two named entities (or high frequency nouns) and verbs (or action nouns) are atomic relations);
4. We score the atomic relations identified in the retrieved documents.[7]

We also extend the list of high frequency nouns with the terms used in the description of the *event* in question. We assumed that all the nouns used in the formulation of the

[7] For more details on atomic relation extraction and their scoring see Chapter 3.

event in question are important constituent parts of the event. Thus, after we identify the ten most frequent nouns used in the IR output, we extend this list with the nouns used in formulation of the *event* in question. After that, we extract all atomic relations mentioned in the set of documents returned by Indri for the question under investigation. We do not evaluate explicitly the quality of the extracted atomic relations. Rather, we use these relations as sentence selection features: we assign scores to all the sentences in the set of retrieved documents according to how many and what atomic relations are covered by each sentences. We then select sentences starting with the one that covers the most atomic relations with the highest scores. We then evaluate the quality of the answer constructed out of the selected sentences. It must be noted that the shallow semantic network was designed to capture the structure of the event described in a document collection. According to the summarization experiment presented in Chapter 3, the sentence selection procedure guided by a shallow semantic network gives the best performance for those document collections that describe events with well-defined set of constituent parts (people, locations, dates). Thus, if the IR output contains a lot of information that is not relevant to the event description, then the shallow semantic network extracted from this document collection contains a substantial amount of noise. Consequently, this network leads the sentence selection procedure astray and the sentences included in the answer do not contain any relevant information. For example, the set of the high frequency nouns that were identified for the question regarding the shut down of the Cernavoda nuclear power plant was the following: *history*, *Jews*, *holocaust*, *gypsies*, *ceremony*, *water*, *minister*, *government*, *war*, *electricity*. Most of these high frequency nouns are coming from the documents about the history of Holocaust is Romania and, obviously, the shallow semantic network constructed according to such elements forced our sentences selection algorithm to add to the answer a substantial amount of irrelevant sentences.

To select answer sentences we use the modified adaptive greedy algorithm described in Chapter 3. as this algorithm showed the best performance for the event-based summarization task. For the QA task, as it was formulated for GALE Year 1 Go/No-Go evaluation, there was no limit on the length of the answer. Thus, as the stopping criteria for the modified adaptive greedy algorithm for the version 1 of our system, we chose the absence of sentences containing important atomic relations. To avoid selecting sentences corresponding to the atomic relations with low scores, we use only those atomic relations where the count for the pair of named entities (or high frequency nouns) is more or equal to two. This parameter was estimated empirically as prior to the Go/No-Go 2006 evaluation there existed no reliable models against which the quality of the answer could be compared.

5.3.1.3 System Version 1: Evaluation

The results for the first experiment were rather poor. For some questions, many relevant sentences were missed, while for other questions many irrelevant sentences were included in the answer. For example, for the question

> "List facts about event [*The shut down of the Cernavoda nuclear power plant*]"

many returned documents described the history of Holocaust in Romania.[8] As the number of documents on Romania and Holocaust exceeded the number of documents on the actual shut down of the Cernavoda nuclear power plant, many of the high frequency nouns are not linked to the event in question. This, in its turn, led the shallow semantic network astray from the structure of the event we wanted to induce. In addition, atomic relations extracted from the relevant documents got lower scores and were used to selected sentences that were added to the second half of the answer. Thus, many of the sentences selected for the answer were irrelevant.

The answers submitted by our system and the other two participating systems were judged manually. The recall, precision and F-measure values manually computed by BAE annotators are presented respectively in Figure 5.4, Figure 5.5, Figure 5.6 and are marked as ***BAE prec***, ***BAE recall*** and ***BAE F-meas***. Note that for question GNG041 our system failed to produce any answer due to the inability of dealing with speech documents and thus, a back-up system was used instead.

In Section 5.2.3 we described the procedure for automatic evaluation of the answer quality using cosine similarity. The automatically computed precision, recall and F-measure

[8] See Section 5.3.1.1 on more detailed analysis of the documents returned for this question.

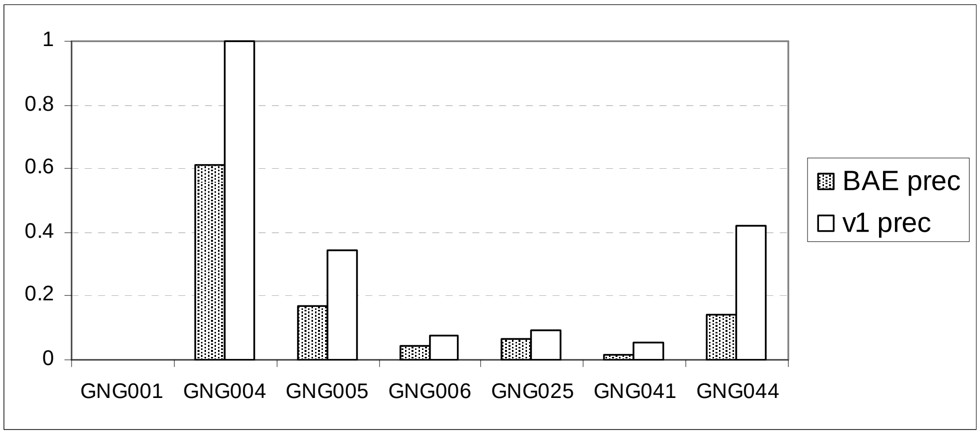


Figure 5.4: Precision scores for the seven Template 1 questions used in Year 1 GALE Go/No-Go formal distillation evaluation (system version 1).

in Figure 5.4, Figure 5.5, Figure 5.6 and are marked as ***v1 prec***, ***v1 recall*** and ***v1 F-meas***. Analyzing the data presented in Figures 5.4 – 5.6 we can see that though the precision, recall and F-measure computed automatically are not equal to the precision, recall and F-measure computed manually, they track the manually computed metrics rather well. Thus, we decided to use our automatic tool to evaluate answers quality for the other two versions of our system. The ability to automatically compute scores for the three versions of our system allowed us to track the development the system.

For several questions, the sentences selected for the answers were drawn from documents that were produced by machine translation (MT) or automatic speech recognition (ASR) systems. As all three participating systems used different systems for MT and ASR computing cosine similarity for ASR and MT answer snippets causes a lot of errors (many relevant sentences are judged as irrelevant). For the Template 1 questions used for our experiments, the answers for questions GNG001 and GNG041 are drawn only from ASR documents.

The output of our initial QA system for Template 1 together with the automatic relevance judgement of the answer sentences is presented in Appendix B, Section B.1.

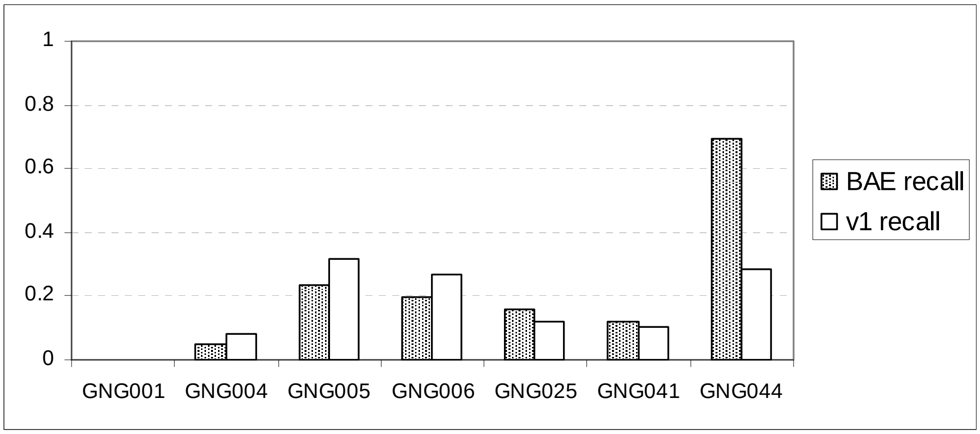


Figure 5.5: Recall scores for the seven Template 1 questions used in Year 1 GALE Go/No-Go formal distillation evaluation (system version 1).

5.3.1.4 System Version 1: Error Analysis

After the Go/No-Go evaluation, we performed error analysis of our system. To address the problems that we discovered through the data analysis we introduced several changes targeting the major error sources. In our second and third experiments, we used two updated versions of the initial system to produce answers to the questions used in July 2006 Go/No-Go formal distillation evaluation and compared these answers to the ones produced in the first experiment. In the rest of this section we describe the set-up of the three versions of the system and compare the results produced by these system versions.

For all three versions of our QA system, the final answer consisted of a set of sentences. The sentence selection process in all versions was guided by shallow semantic network. However, the parameters used for constructing shallow semantic network and for scoring atomic relations were different for different versions. In the remaining of this section we first describe the Information Retrieval system that we used to get document collections for our QA system. Then, we describe the three versions of our QA system targeting GALE Template 1 questions, where the first version of the system is the one that was used for July 2006 Go/No-Go evaluation. We also describe the automatic evaluation tool that we used to track the development of our system through its tree versions. Finally, we discuss the comparative evaluation of the performance of these three systems.

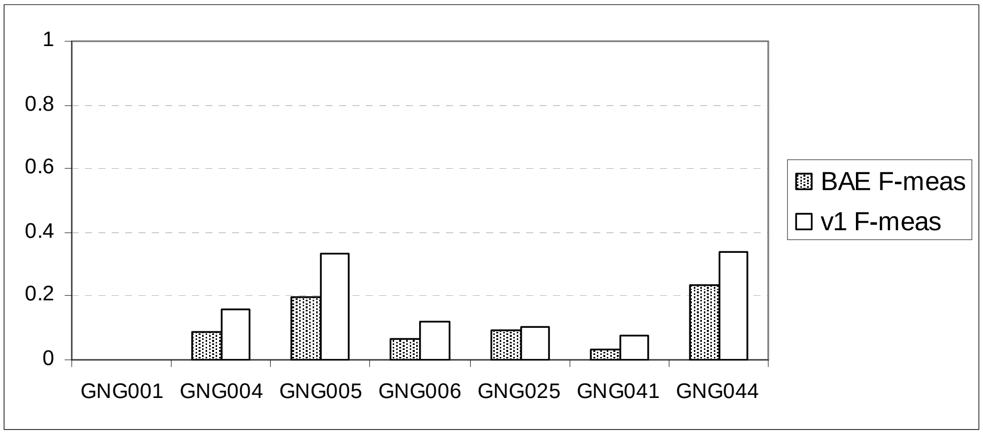


Figure 5.6: F-value scores for the seven Template 1 questions used in Year 1 GALE Go/No-Go formal distillation evaluation (system version 1).

We analyzed the answers that we submitted for the June 2006 Go/No-Go evaluation (Appendix B, Section B.1) and identified two major sources of errors:

Error Source 1: The run of Version 1 of our system showed that the IR output that we use as an input for our answer selection procedure is likely to contain a substantial amount of irrelevant documents. Thus, the answer should be created according to the assumption that it is likely that IR output contains substantial amount of noise. Obviously, building a shallow semantic network treating all the documents as equally relevant caused a lot of errors. Thus, addressing the issue of noisy IR output was essential to improve the precision of our system.

Error Source 2: The modified adaptive greedy algorithm was developed to pack as much of unique important information facts in a short text as possible. In the GALE evaluation, however, no limit was posed on the answer length. Rather, it was desirable to include in the answer as many relevant sentences as possible. A subsequent module identifies redundancy for the response (Thadani 2007). Sentences that were included in the answer were either answer snippets or sentences redundant to the answer snippets. Clearly, using the modified adaptive greedy algorithm developed for efficient information packaging hurt the recall of our QA module.

To improve the quality of our answers we decided to combat the error sources one at a time. For Version 2 of our system we developed a two-step information retrieval procedure.

We show that analyzing first high-precision documents and then propagating this analysis to a larger IR output improves the quality the answers produced by our QA system. For Version 3 of our system we modified the answer selection procedure to address the erroneous use of information packaging procedure.

5.3.2 System Version 2

For Version 1 of our system, we assume that all the documents retrieved by Indri are equally relevant. This assumption makes it possible for our system to draw answer sentences from irrelevant documents. The analysis of answers produced by this system shows that treating all the returned documents equally might cause the addition of irrelevant text snippets to the answer. To address the issue of irrelevant answer snippets due to the noise in the IR output we designed and implemented a two-step IR procedure where the first step retrieves documents of high precision while the second step targets high recall in IR output.

5.3.2.1 System version 2: Information Retrieval Stage

In our second experiment, we implemented a two-step IR procedure, where in the first step only highly precise documents were analyzed, while in the second step our goal is to retrieve as many relevant documents as possible. A similar approach for addressing separately precision and recall for both IR documents and the answer sentences that are selected from these documents is used in (Schiffman et al. 2007).

IR first step: Submit a query requesting documents which include *a*ll the terms used in the description of the *event* in question. This is a strict boolean query. Given such queries, for most questions Indri returns a small number of relevant documents. For some of the queries no documents are returned for the submitted strict query. In this case, as high-precision documents we analyze the top three documents returned by the probabilistic Indri search procedure for Template 1 questions for the input question. We use the set of high-precision documents to extract the list of high frequency nouns.

IR second step: Get the same set of documents as was used in the first version of the system. We do not recalculate the set of high frequency nouns for this stage and thus, we avoid high frequency nouns from irrelevant documents to be included in this list. Use the list of high frequency nouns constructed at the high-precision IR step, extract all the atomic relations and create a shallow semantic network for the current set of IR documents.

5.3.2.2 System Version 2: Sentence Selection Stage

For this version of the system we kept the answer sentence selection procedure that was used within the first version of the system. We, first selected sentences from the high-precision documents using the shallow semantic network built for those documents. Then, we extended our shallow semantic network for the set of documents retrieved at the second IR step, and, following the sentence selection procedure used in Version 1 of our system selected those sentences that cover parts of the shallow semantic network and at the same time are not already added to the answer due to the presence in high-precision documents.

5.3.2.3 System version 2: Evaluation

To evaluate the improvement of our QA system after introduction of two-step IR we use the automatic evaluation procedure described in Section 5.2.3. The quality of the answers submitted by this version of our system are compared against the quality of the answers submitted by the initial version of our system. Precision, recall and F-measure are presented respectively in Figures 5.7, 5.8, and 5.9. The results for Version 1 of our system are marked as ***v1 prec***, ***v1 recall*** and ***v1 F-meas***. The results for the current version of our system (Version 2) are marked as ***v2 prec***, ***v2 recall*** and ***v2 F-meas***. The actual answers of Version 2 for Template 1 together with the automatic relevance judgements of the answer sentences are presented in Appendix B, Section B.2.

The goal of two-step IR was to improve the shallow semantic network extraction procedure, so that it captured the structure of the event in question, rather than the structure of the document collection retrieved to the corresponding question. By improving the quality of the extracted semantic network we minimize the chances of irrelevant sentences to be included in the answer. According to the results presented in Figures 5.7, 5.8, and 5.9, we achieved our goal, as the introduction of the two-step IR procedure caused a significant improvement of the answers precision. According to Figure 5.7, we get higher precision for three out of seven questions. It must be noted that for GNG001, the precision of our answer

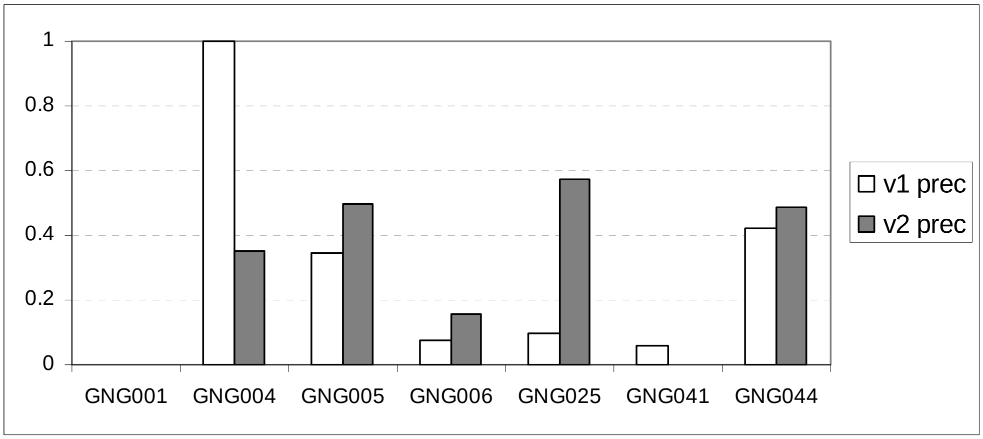


Figure 5.7: Precision scores for the seven Template 1 questions used in Year 1 GALE Go/No-Go formal distillation evaluation (system version 2).

is higher as well. All the answer sentences, however, come from ASR (Automatic Speech Recognition) documents and thus, are not captured as relevant by our automatic procedure based on cosine similarity scores. It is obvious, though, that sentences one, two and three for the GNG001 question presented in Appendix B, Section B.2 are relevant:

1. his office and one witness suggested whittington was at fault because he failed to alert mr. cheney that he was rejoining the hunting party after searching for a down to earth

2. vice president dick cheney under pressure goes public tonight about shooting his hunting companion

3. when vice president dick cheney accidentally shot a member of his hunting party

Another case where the precision for the current version of the system is lower than the precision of the first version of the system is GNG004. It must be noted, though, the first time around we returned only two sentences both of which were judged as relevant, while for this version of the system we returned 17 sentences, six of which are judged as relevant.

According to Figures 5.7, 5.8, and 5.9, it can be concluded that introduction of two-step IR improved the overall performance of our system. It must be noted, that though the major goal of two-step IR procedure was to get higher precision scores, it also helped recall scores and, consequently, F-measure.

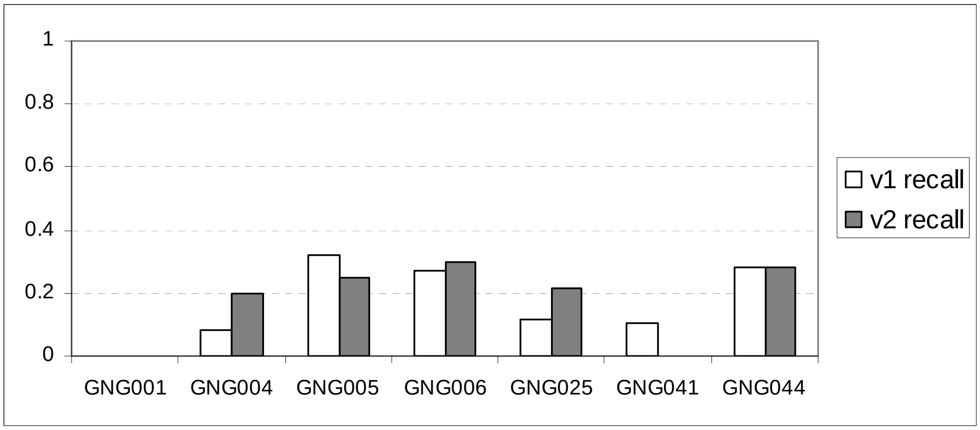


Figure 5.8: Recall scores for the seven Template 1 questions used in Year 1 GALE Go/No-Go formal distillation evaluation (system version 2).

5.3.3 System Version 3

In the third version of our system, we address the problem of low recall of our system. As noted earlier, we believe that this problem was caused by the use of the information packaging procedure in Version 1 of our QA system. This information packaging procedure was developed specifically for the summarization task where it is important to put as much important information in a text of fixed length. For the GALE evaluation, there exists no limit on length of the output. Thus, for the third version of our system, we eliminated from the answer selection procedure the information packaging model. It must be noted, though, that the Columbia QA system evaluated within the DARPA GALE evaluation framework had a separate module the main goal of which was to group together those sentences that had overlap in the information covered by them (Thadani 2007).

5.3.3.1 System Version 3: Information Retrieval Stage

The two-step IR procedure tested within the second version of the system caused improvement in both precision and recall scores for the answers produced by our QA system in comparison to the answers produced by QA system with IR procedure where all the documents are treated equally. Thus, for the third version of our system we used the same two-step IR procedure that we used for Version 2 of the system.

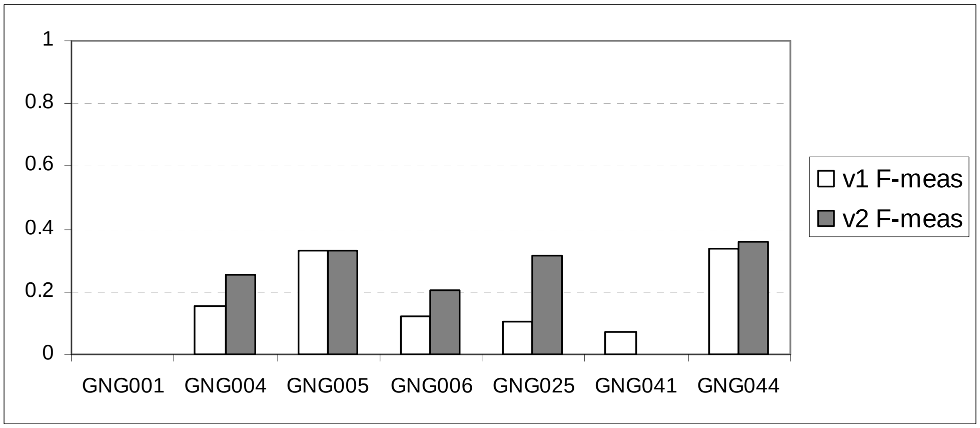


Figure 5.9: F-value scores for the seven Template 1 questions used in Year 1 GALE Go/No-Go formal distillation evaluation (system version 2).

5.3.3.2 System Version 3: Sentence Selection Stage

In the third version of our system we address the **Error Source 2** identified in Section 5.3.1.4. As noted, a lack of relevant sentences in the answer could have been caused by the information packaging methodology that we developed for the summarization task and used as part of the answer sentence selection procedure in Versions 1 and 2 of our QA system. Thus, in Version 3 we excluded the information packaging from the answer sentence selection procedure. Thus, we do not use the modified adaptive greedy algorithm for answer sentence selection; instead we select all the sentences containing important atomic relations. This sentence selection procedure is applied for each IR step.

Sentence selection for high-precision IR The sentences selected for this set of documents are placed at the very beginning of the answer. The sentence selection process ends when all the sentences containing important atomic relations have been selected. The atomic relations are considered important if the pair of named entities or high frequency nouns appears in the high precision documents two times or more; and the atomic relation connector links these named entities or high frequency nouns in the high precision documents two times or more. The rest of the atomic relations are disregarded in further analysis as they are likely to capture noise rather than important relations. Then, each sentence in the set of high precision documents is scored according to how many important atomic relations

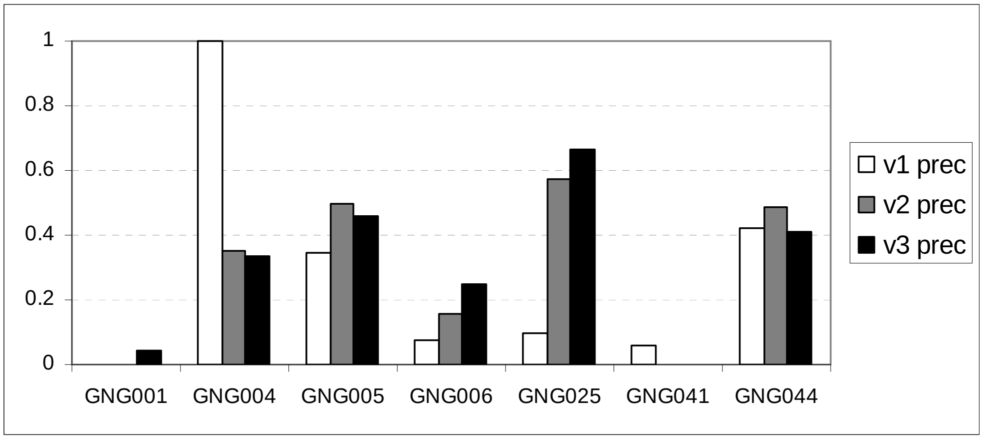


Figure 5.10: Precision scores for the seven Template 1 questions used in Year 1 GALE Go/No-Go formal distillation evaluation (system version 3).

5.3.3.3 System Version 3: Evaluation

To evaluate the improvement of our QA system after the introduction of two-step IR and removal of the information packaging procedure, we use the automatic evaluation procedure described in Section 5.2.3. The quality of the answers submitted by Version 3 of our system is compared against the quality of the answers submitted by Versions 1 and 2. Precision, recall and F-measure are presented in Figure 5.10, 5.11, and 5.12 respectively. The results for Version 1 are marked as ***v1 prec***, ***v1 recall*** and ***v1 F-meas***. The results for Version 2 are marked as ***v2 prec***, ***v2 recall*** and ***v2 F-meas***. The results of the current version of our

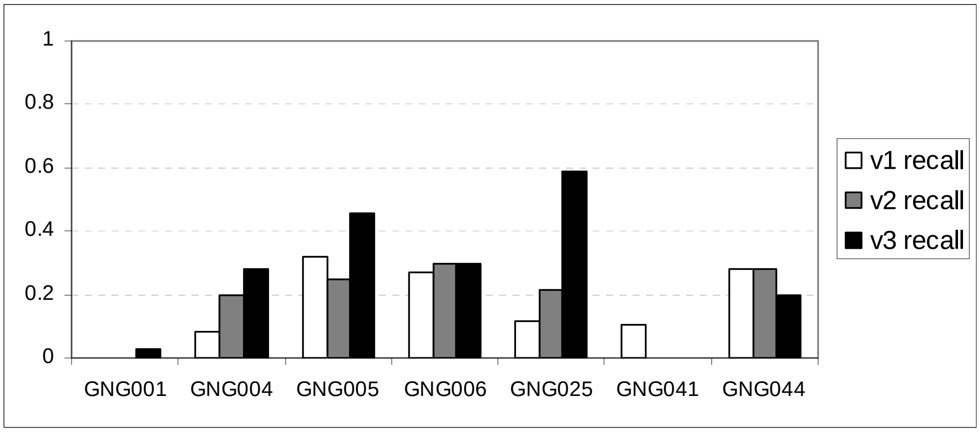


Figure 5.11: Recall scores for the seven Template 1 questions used in Year 1 GALE Go/No-Go formal distillation evaluation (system version 3).

system (Version 3) are marked as ***v3 prec***, ***v3 recall*** and ***v3 F-meas***. The actual answers of the third version of QA system for Template 1 together with the automatic relevance judgement of the answer sentences are presented in Appendix B, Section B.3.

According to the results presented in Figures 5.10, 5.11, and 5.12, the two changes introduced after the error analysis of the initial system improved the quality of the answers produced by our system to GALE Template 1 questions.

We also compared the best performing version of our system against the other two systems that participated in the GALE challenge, namely, BBN and IBM systems. The comparisons of precision, recall and F-measure are presented in Figure 5.13, 5.14, and 5.15 respectively. According to the evaluation procedure design, the set of snippets included into the answers of system version 3 were not included in the answer models. The answer snippets that are included in the BBN and IBM answers, on the contrary, were included in the answer models. It has been noted that the answers for questions GNG001 and GNG041 are drawn only from ASR documents. The cosine similarity comparison does not work on ASR text as consistently as it works on plain English text. Thus, the results of our system for questions GNG001 and GNG041 look poor. For the rest of the questions the performance of our system is comparable to the performance of the other two systems.

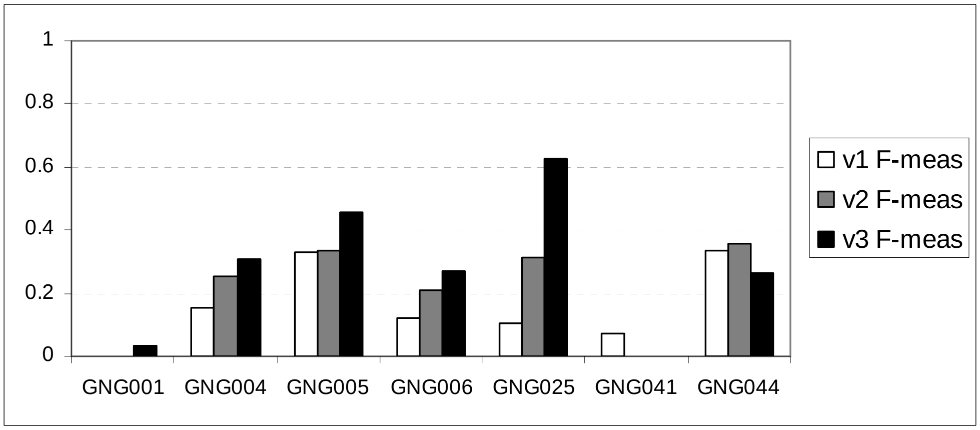


Figure 5.12: F-value scores for the seven Template 1 questions used in Year 1 GALE Go/No-Go formal distillation evaluation (system version 3).

5.3.4 System Limitations

Currently, we do not take into consideration the semantics of the event in question. Rather, we use the event description as a set of lexemes for query generation, and a subset of high frequency nouns for a shallow semantic network induction. Knowing the semantics of the event in question can be beneficial for any stage of a QA pipeline. For example, the semantic information about the event in question can be used for deducing terms that a highly related to the event and can be used for query expansion. Also, this information can be used to predict the structure of the event and thus, improve the quality of the answer. Both query expansion and targeted answer selection knowing the semantics of the event in question were implemented within our system to answer Template 8 questions about prosecutions.

5.4 Answering *Template 8* Questions (Prong 2)

For the questions corresponding to GALE Template 8 (Describe the prosecution of [*person*] for [*crime*].) the semantics of the event that should be covered in the answer is explicitly stated in the question. Thus, we can use this information to guide both the information retrieval and answer selection processes.

Two specialized answer generation modules for answering Template 8 questions were implemented within the Columbia event-related open-ended QA system. One module is

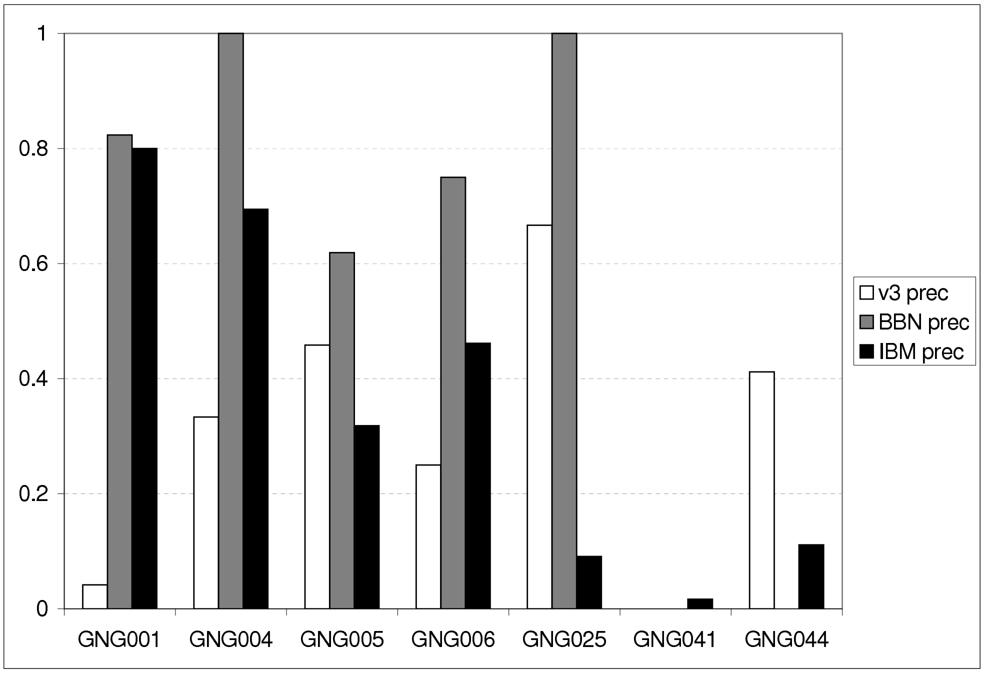


Figure 5.13: Precision scores for the seven Template 1 questions used in Year 1 GALE Go/No-Go formal distillation evaluation (system version 3, BBN, and IBM systems).

based on the automatically created domain template for the *prosecution* domain; the other module is based on the set of ACE *justice* event subtypes, where these subtypes can be considered as manually defined *prosecution* domain template slots. In this section, we describe the former system for answering Template 8 questions. The detailed explanation of the latter system can be found in (Schiffman et al. 2007).

5.4.1 Domain Template Generation Stage

To create a domain template for the *prosecution* domain we used the procedure described in Chapter 4 with a slight modification at the stage of creating document collections corresponding to different instances of the *prosecution* domain. To create document collections corresponding to the *prosecution* domain instances we used the Indri search engine and submitted a query consisting of only one query term *prosecution*. We requested 500 documents to be returned. Afterwards, we clustered the retrieved documents setting a threshold to be rather high. We retained for further consideration only those document clusters that contained seven or more documents. There were six clusters that satisfied this criteria. Using

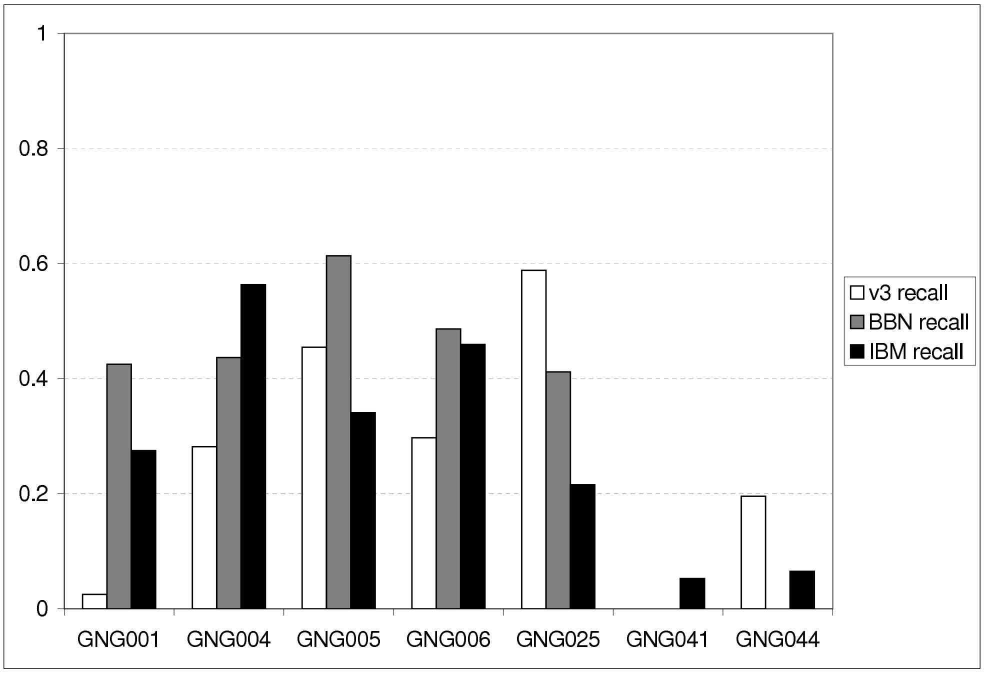


Figure 5.14: Recall scores for the seven Template 1 questions used in Year 1 GALE Go/No-Go formal distillation evaluation (system version 3, BBN, and IBM systems).

these six document clusters we created the *prosecution* domain template following exactly the procedure described in Chapter 4. For example, here are the top ten verbs that were identified for the *prosecution* domain:

> *including*, *told*, *signed*, *expected*, *accused*,
> *make*, *prosecute*, *charged*, *committed*, *continue*.

5.4.2 Information Retrieval Stage

To retrieve documents corresponding to Template 8 questions we used the same IR module that was developed by Schiffman *at el* (2007). The queries corresponding to the questions were expanded using pre-computed terms that are associated with the prosecution event. These terms were determined on training data corresponding to descriptions of prosecutions. Also, the query was expended using the terms that corresponded to different references to the person mentioned in the question. For example, the final query for the question regarding Saddam Hussein's prosecution for crimes against humanity is presented in Figure 5.16.

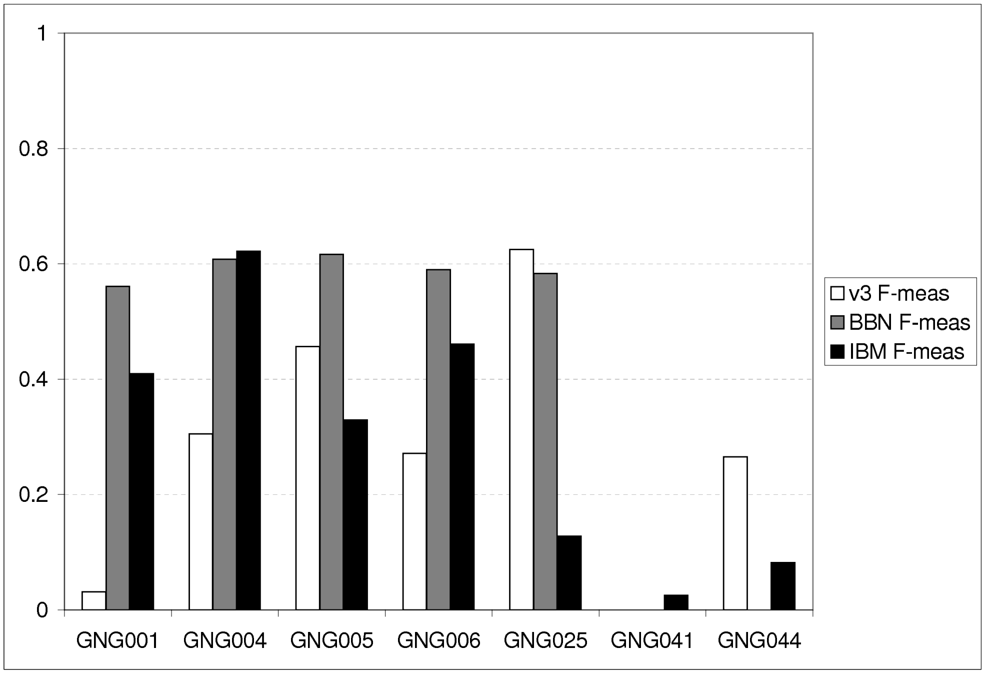


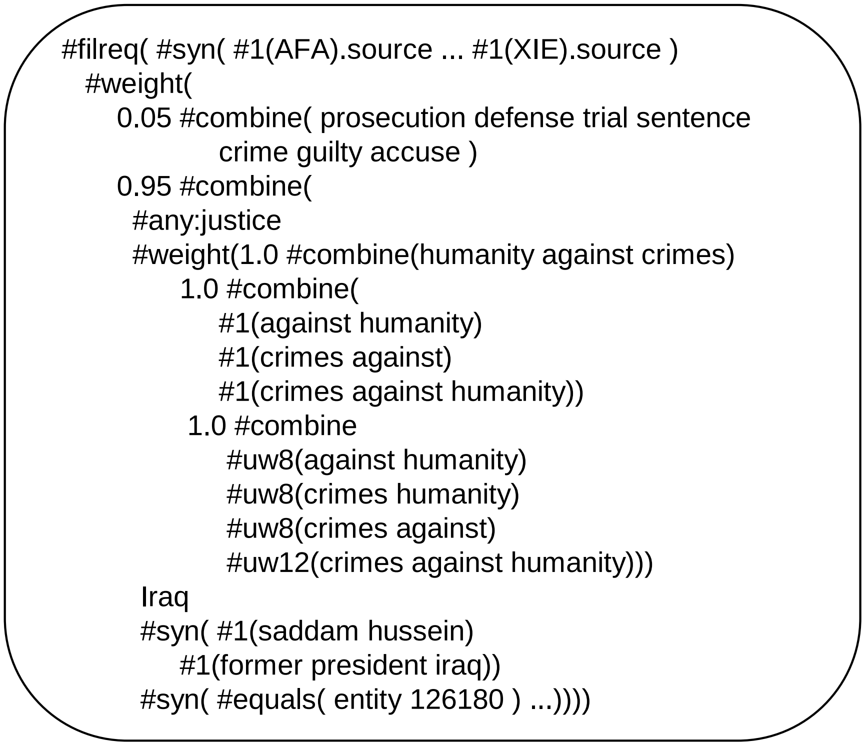
Figure 5.15: F-value scores for the seven Template 1 questions used in Year 1 GALE Go/No-Go formal distillation evaluation (system version 3, BBN, and IBM systems).

For our experiments, we requested ten documents from the Indri search engine, this is the number of documents requested by the system based on the ACE *justice* event subtypes that was the primary system for answering Template 8 question during the Year 1 GALE Go/No-Go evaluation.

5.4.3 Sentence Selection Stage

For sentence selection we used two approaches: one based on the automatically created *prosecution* domain template, and the other one based on shallow semantic network.

For the former approach we used the procedure described in Chapter 4 with one small modification: when searching for a text snippet to cover a specific template slots we did not analyze simple clauses; rather, we searched only complete sentences. For the latter approach we used the portion of the procedure described in Section 5.3.3.2 that corresponds to the answer selection for the set of high-precision documents. We used the answers created according to the shallow semantic network for comparative evaluation.



```
#filreq( #syn( #1(AFA).source ... #1(XIE).source )
  #weight(
     0.05 #combine( prosecution defense trial sentence
             crime guilty accuse )
     0.95 #combine(
      #any:justice
      #weight(1.0 #combine(humanity against crimes)
          1.0 #combine(
             #1(against humanity)
             #1(crimes against)
             #1(crimes against humanity))
          1.0 #combine
             #uw8(against humanity)
             #uw8(crimes humanity)
             #uw8(crimes against)
             #uw12(crimes against humanity)))
      Iraq
      #syn( #1(saddam hussein)
          #1(former president iraq))
      #syn( #equals( entity 126180 ) ...))))
```

Figure 5.16: The Indri query for the Template 8 question DESCRIBE THE PROSECUTION OF [*Saddam Hussein*] FOR [*crimes against humanity*].

5.4.4 Evaluation

To evaluate the quality of the answers created for Template 8 questions we used the automatic procedure described in Section 5.2.3. The answers to the Go/No-Go Year 1 Template 8 questions together with the automatically generated relevance labels are presented in Appendix C, Section C.1 contains a set of answers that were created using the automatically induced *prosecution* domain template and the sentence selection procedure described in Chapter 4, while Section C.2 contains a set of answers that were created using a shallow semantic network.

Figures 5.17, 5.18, and 5.19 present the automatically computed prosecution, recall and F-measure values. The ***CU*** results correspond to the answers submitted for the Go/No-Go evaluation. These answers were created using ACE *justice* event subtypes (Schiffman et al. 2007) and a generalized version of the DefScriber system (Blair-Goldensohn, McK-

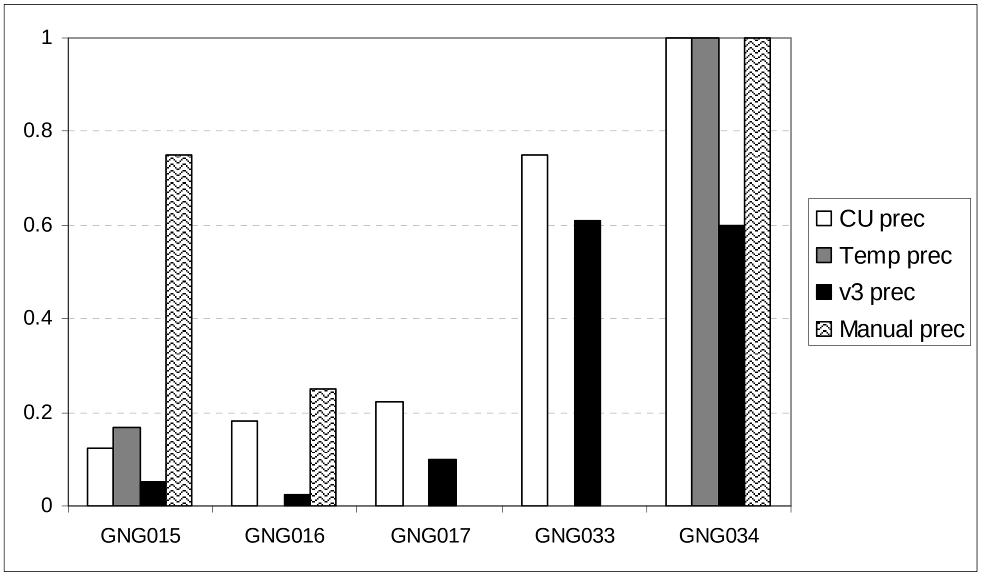


Figure 5.17: Precision scores for the five Template 8 questions used in Year 1 GALE Go/No-Go formal distillation evaluation.

eown & Schlaikjer 2004). The ***Templ*** results correspond to the answers selected using the automatically induced *prosecution* domain template. It must be noted, that for the GNG033 question no answer sentence was selected using the *prosecution* domain template. The ***v3*** results correspond to the answers selected using a shallow semantic network.

We also asked Barry Schiffman to judge manually the answers produced according to the automatically created domain template. He previously ran several evaluation experiments for questions corresponding to GALE Template 8 and studied the results of these experiments. Thus, we believe that his judgement was highly reliable. The precision results corresponding to manual annotation of the results are marked in Figure 5.17 as ***Manual prec***. The difference between the manual and automatic evaluation is significant for the GNG015 and GNG016 questions. Thus, for the GNG015 question, according to the manual evaluation three out the twelve answer sentences, namely 5, 6, and 10, are judged as irrelevant, while according to the automatic evaluation ten out of the twelve answer sentences are judged as irrelevant.[9] For the GNG016 question, according to the manual evaluation out of the four answer sentences, one sentence (namely, sentences 3) is relevant, while according

[9] See Appendix C, Section C.1.

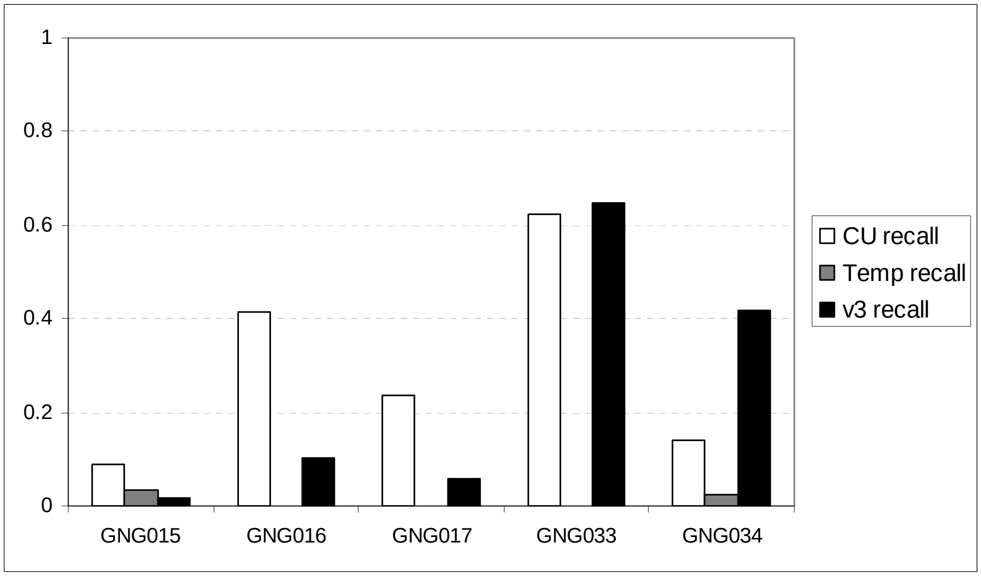


Figure 5.18: Recall scores for the five Template 8 questions used in Year 1 GALE Go/No-Go formal distillation evaluation.

to the automatic evaluation none of the four answer sentences is relevant. The significant difference in scores can be attributed to the fact that many answer sentences were drawn from the documents automatically translated into English from Arabic and Chinese. For such sentences, the cosine similarity metric used in our automatic evaluation procedure is not reliable and is likely to mark relevant sentences as irrelevant.

Figure 5.20 contains information on how many answer sentences were produced by the three systems evaluated in Figures 5.17, 5.18, and 5.19. According these four figures, it can be concluded that the answers produced by the system based on the automatically induced *prosecution* domain template are very short and the precision of these answers is rather high; the recall of these answers, however, is quite low. This result was predictable, as domain templates are designed to capture general domain information but at the same time, each domain instance has its own peculiarities that are important for that particular instance but are not typical for the domain in general and thus, are not covered by a domain template.

Also, with respect to the quality of the answers produced by the system based on automatically induced *prosecution* domain templates, we would like to emphasize the fact that

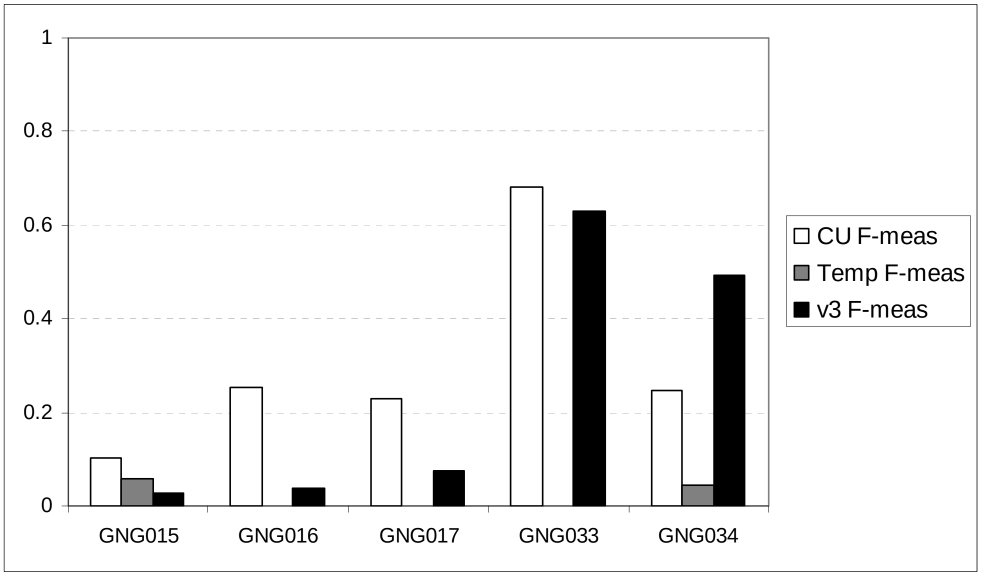


Figure 5.19: F-value scores for the five Template 8 questions used in Year 1 GALE Go/No-Go formal distillation evaluation.

the document collections corresponding to the domain instances were created automatically through clustering. Thus, we believe, that the result high-precision answers that were produced using less than reliable data are very promising. For example, if the domain of the event in question can be induced on-the-fly, we can create a completely automatic pipeline for generating high-precision answers focused on the information relevant to the domain of the event in question.

5.5 Conclusions and Future Work

In this Chapter we described three versions of the QA system we developed for answering event-related questions. We evaluated the answers produced by the three versions of our system for GALE Template 1 questions used for July 2006 Go/No-Go distillation evaluation. We designed and implemented a two-step IR procedure that targets separately the precision and recall portions of the answer selection process. We showed that using 2-step IR we can construct better shallow semantic network, which, in its turn, leads towards better answer quality. We developed a procedure that interleaves information retrieval and answer

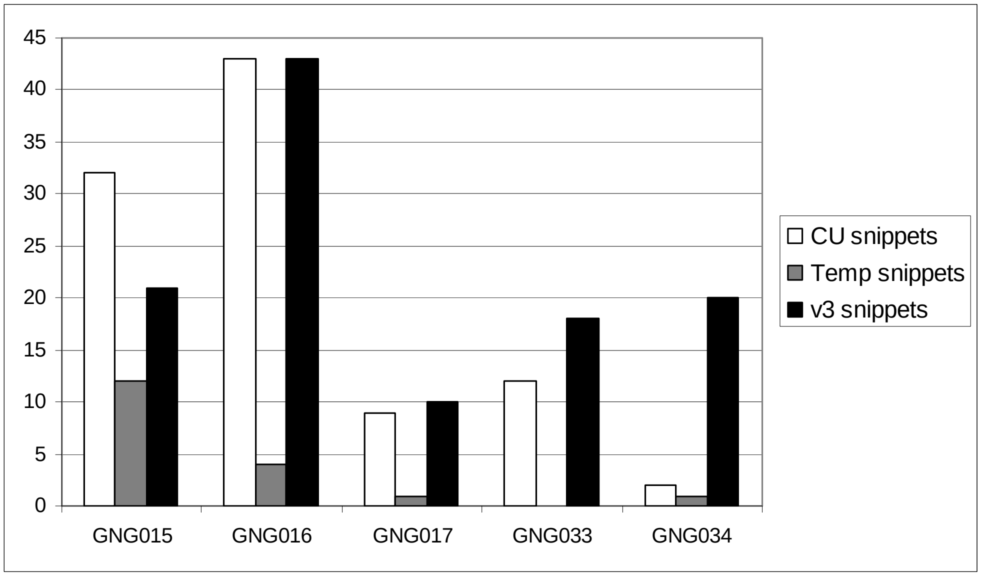


Figure 5.20: Number of answer sentences produced for the five Template 8 questions used in Year 1 GALE Go/No-Go formal distillation evaluation.

selection stages targeting a high-precision and yet containing as much as possible of relevant information answer. We also demonstrated that the information packaging procedure that was developed for a summarization task can decrease the quality of a QA system, if a QA system does not have a length limit for its answers.

Currently, we do not use any semantic information that can be deduced from the description of the *event* in question. Rather, we treat this description as a bag-of-words and use it to get high-precision documents and to extend the list of high frequency nouns for the shallow semantic network. We believe, that using the semantics of the *event* in question is the necessary next step in the development of our QA system. Knowing semantic information encoded in the question can lead to improvements of our system on different levels. First, semantic information can be used to assign the *event* in question to a particular domain for which a domain template exists. It has been shown that if the domain of the *event* in question can be identified, the presence of a good quality domain relation in a sentence is a strong indicator that this sentence should be included in the answer. This approach to answer sentence selection has been used as part of the Columbia GALE distillation system within the QA module that targeted *Template 8* questions about prosecutions (Schiffman et al. 2007, Hakkani-Tür et al. 2007). *Prosecution* domain relations used by the both systems were manually identified as subtypes of *justice* ACE event. In the experiment we describe in Chapter 4, we demonstrate that automatically identified domain relations are also reliable sentence selection features for questions about domain events. On the other hand, if the *event* in question cannot be assigned to a domain from the list of domains for which domain templates exist, semantic information encoded in the question can be used to identify the event constituent parts and thus, lead towards improved shallow semantic network.

Chapter 6

Biography as a Special Case of Domain

When you read a biography
remember that the truth is
never fit for publication.

George Bernard Shaw

In Chapter 4 we described the procedure we developed for automatic domain template induction. This procedure does not require any preliminary knowledge of the domain. The core part of the procedure is based on cross-analysis of different instances of the same domain. It was suggested that using the described domain template induction procedure it is possible to automatically build a domain hierarchy.

In this chapter, we discuss a special case of a domain, a ***biography*** domain. Biography creation requires the identification of important activities in the life of the individual in question. While there are activities such as birth and death that apply to everyone, most of the other activities tend to be occupation-specific. Hence, occupation gives important clues as to which activities should be included in the biography. In this chapter, we present techniques for automatically identifying which *important* activities apply to the general population, which ones are occupation-specific, and which ones are person-specific. We use the extracted information as features for a multi-class SVM classifier, which is then used to automatically identify the occupation of a previously unseen individual. We also show that automatically extracted human activities can be used not only for the classification task but for clustering as well. We show that, given the correct number of classes, people belonging to the same occupation are clustered together. At the same time, clustering people into a smaller number of classes allows grouping of practitioners within those applications who share a considerable number of occupation-specific activities. For example, clustering practitioners of various occupations into two clusters provides us with grouping mathematicians, physicists and explorers in one cluster and artists, composers, dancers, singers, writers in the other cluster. This separation of people into two clusters roughly corresponds to the division of occupations and practitioners of these occupations into arts and sciences. Thus, analyzing descriptions of people belonging to various occupations, we can build a hierarchy of occupations. Each node of this hierarchy includes not only a list of related occupations, but also practitioners of these occupations as well as a list of activities typical to people belonging to these occupations. The work described in this chapter is a joint work with Dr. Prager from IBM T.J. Watson Research Center.

6.1 Related Work

Recently there has been increased interest in creating systems which combine summarization and QA and can give long answers to definition, biography, opinion and other question types. Such systems use general summarization techniques and at the same time, take advantage of the fact that the selected information is not *generally* important but should be targeted towards answering the user's request. For example, the systems participating in DUC 2004 create summaries which could be used as answers to the question "Who is X?". These systems use a wide variety of techniques. Blair-Goldensohn et al. (2004) treat "Who is X?" as a definition question and use the DefScriber system (Blair-Goldensohn, McKeown & Schlaikjer 2004) to create focused summaries that would correspond to biographies. The DefScriber system uses a combination of goal-driven and data-driven techniques. First, *definitional predicates* are used to select information suited for a definition (e.g., genus-species information), and the rest of the answer is shaped according to the themes found in the input document collections. To make biography text more coherent, DefScriber uses a system for reference rewriting (Nenkova & McKeown 2003). Biryukov et al. (2005) use Topic Signatures (Lin & Hovy 2000) constructed around the person's name, as Topic Signature generation exploits the natural tendency of the semantically related words to co-occur more often than by chance in the same context. Zhou et al. (2004) present a system created specifically for the biography generation task. It uses nine features which are likely to be used in biography texts: bio (biographical facts), fame, personality, social, education, nationality, scandal, personal, work. Using manual annotation of 130 biographies they learn the textual patterns corresponding to these nine features. In this chapter, we suggest an algorithm that can be used for unsupervised learning of the features similar to the nine manually defined biography-related features used in (Zhou et al. 2004).

Biographical information can be used to answer not only "Who is X?" questions. Prager et al. (2004) use biographical information within their QA-by-Dossier-with-Constraints system, which checks whether the possible answer satisfies the constraints for the person about whom the question is asked. One of such natural constraints for artists, composers and writers is that all their works are produced in the span of time between the dates of birth and death. For example, for the question "When did Leonardo da Vinci paint the Mona Lisa?" five candidate answers with their initial confidence scores are presented in Table 6.1.[1] It must be noted that, the initial scores are computed without taking into consideration the biographical information. The correct answer **1503** is scored fourth. However, using Leonardo da Vinci's biographical information such as, he was born in 1452 and died in 1519, makes only two answers from Table 6.1 possible. By choosing out of these two possible answers the one with the highest score we get the correct answer to the questions "When did Leonardo da Vinci paint the Mona Lisa?"

To extract biographical facts it is useful to understand the nature of different human activities. Biographical activities such as birth, death, and living somewhere are applicable to all people, while each occupation is associated with its own set of activities.

Knowing biographical activities can be helpful for IR applications as a significant percentage of search queries seem to contain references to people. Guha *et al* (2003) describe search over a semantic web where "the semantic web is not a web of documents, but a web of

[1] This example is taken from (Prager, Chu-Carroll & Czuba 2004).

	Score	Painting Date
1	.64	2000
2	.43	1988
3	.34	1911
4	.31	1503
5	.30	1490

Table 6.1: Answers for "When did Leonardo da Vinci paint the Mona Lisa?"

relations between resources denoting real world objects." They show an advantage of having real world objects linked to each other. For example, when asked about ⟨**Yo-Yo Ma**⟩ semantic search outputs a network of objects. This network encodes information that Yo-Yo Ma is a ⟨**musician**⟩, he is connected to ⟨**Paris, France**⟩ by the link ⟨**birth place**⟩, to ⟨**10/07/55**⟩ by the link ⟨**birth date**⟩, etc. Figures 6.1 shows a small chunk of the Semantic Web corresponding to the cellist Yo-Yo Ma.[2]

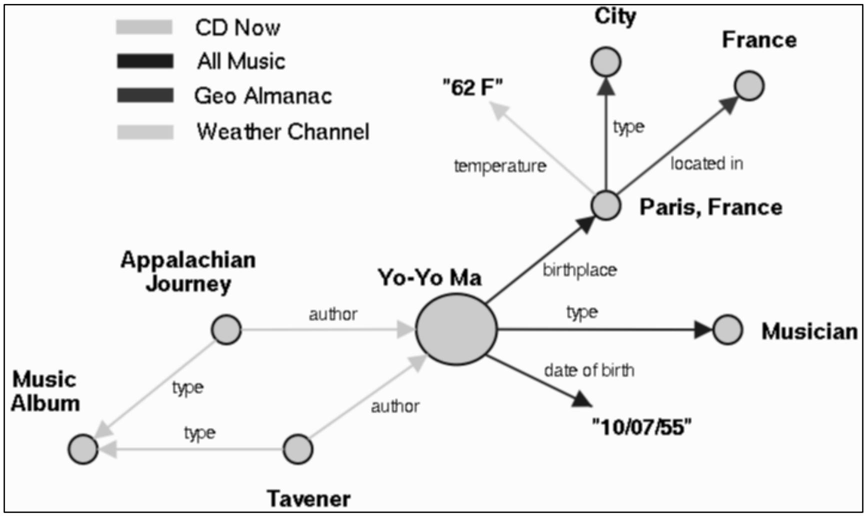


Figure 6.1: A segment of the Semantic Web pertaining to Yo-Yo Ma.

In this chapter we show how *atomic relations* described in Chapter 3 and statistical algorithms based on Markov Chains, can be used for unsupervised extraction of general bi-

[2] This Figure is taken from (Guha, McCool & Miller 2003).

ographical activities (the activities typical for most people irrespective of their occupation), and activities typical for people of a particular occupation.

Biographies vary greatly in length, genre and the presented information they contain. Biographies of the same person in encyclopedias and yellow press magazines might contain different information. Encyclopedic biographies usually contain dates of birth and death, the most important achievements of people, while yellow press magazines tend to describe less important, usually scandalous, facts of someone's life.

Thus, before creating a biography it is important to specify the user's information needs. These specifications define the type of information that should be included in biographies. In this chapter, we are interested in learning what activities are typical for people of different occupations and could be used for encyclopedic biography generation. Thus, we assume that most biographies can be broken into three main parts: biographical facts (the person's place and date of birth, where the person lived); activities typically associated with the person's occupation (e.g., singers sing, explorers travel around the globe to study new lands, artists create paintings); and person-specific activities that make this person unique and distinguish her from other people and/or practitioner of the same occupation. Existing research shows that knowing the persons's occupation is helpful for detecting information which should be used in the biography (Schiffman, Mani & Concepcion 2001, Duboue & McKeown 2003).

The rest of the chapter is structured as following. First, we describe our effort on creating a balanced corpus containing practitioners of a set of occupations. We then describe the procedure we designed and implemented for unsupervised learning of biographical activities used in descriptions of people. We show how random walk theory can be used to rank extracted activities in the order of their importance for a particular occupation. We then demonstrate how this ranking can be used to divide the activities extracted from descriptions of people belonging to the same occupation into three classes: general biographical, occupation-related and person-specific activities. We used occupation-related activities as classification features and evaluate a classifier that assigns a person to a particular occupation based on the occupation-related activities used in the description of this person. Finally, we show that, the activities extracted from people's descriptions can be used not only for classification but for clustering as well. We demonstrate another approach for separating general biographical facts from occupation-related activities and show how clustering can be used for creating hierarchies of domains.

6.2 Unsupervised Learning of Activities Typical for Different Occupations

6.2.1 Data

We had to create our own set of people belonging to various occupations as we were aware of no set diverse enough to analyze activities of people belonging to different occupations. We could not use the list of people whose biographies were created for DUC 2004 task as the input documents for this task were contemporary newswire articles and more than a half of the 50 people used there were politicians.

We therefore collected our own corpus of biographies. We chose 10 occupations and 20 practitioners of each occupation. We understood that 10 occupations would not cover every person mentioned in a news corpus, but that was not critical to our study. We wanted to examine a set that included a number of diverse occupations, along with some that were clearly similar if not directly overlapping.

As described later, we sought documents for each chosen person. Since no documents were found for some of the individuals, these people were eliminated from the experiments; 189 survived. We ended up with the following sets of collections:

For our experiments we use 189 text collections. Documents in each of these collections describe a person or at least mention the name of this person. Each person belongs to one of the ten pre-defined occupations. Thus, we create document collections about:

- 20 artists;
- 18 athletes;
- 20 composers;
- 15 dancers;
- 17 explorers;

- 20 mathematicians;
- 19 physicists;
- 20 politicians;
- 20 singers;
- 20 writers

.

We found that for some occupations human annotators agree upon its representatives however different they are. For example, no matter to school an artist belongs (impressionism, surrealism) he/she is usually addressed as an *artist*. The situation with *politicians* is different. They are often referred to not as politicians, but according to their political parties (e.g., democrat, republican, labourist) or with the help of the post held (e.g., president, prime-minister). Choosing an appropriate occupation title becomes crucial at the document retrieval stage as this title is used to query the search engine.

Our goals for the occupation list are that it satisfies the following criteria:

- it is diverse and covers a substantial variety of occupations from arts, sciences and other aspects of human activities;
- it contains some occupations which are closely related between each other and might be later merged into one superclass, for example, mathematicians and physicists;
- it contains occupations that are very different and it is almost impossible to specify activities which are routinely performed in two occupations, for example, singers and explorers.

To get the lists of people belonging to each particular occupation we use WordNet 2.0 (Fellbaum 1998) (e.g., hyponyms for *composer* contain a list of composers). We also use "Google Sets" interface,[3] it was previously successfully used to find people belonging to the same occupation (Prager et al. 2004).

[3] `http://labs.google.com/sets`

We retrieve documents from four corpora: AQUAINT, TREC, part of World Book and part of Encyclopedia Britannica. For document retrieval we use IBM's JuruXML search engine (Carmel, Amitay, Hersovici, Maarek, Petruschka & Soffer 2001) which allows one to index terms along with any associated named entity class labels. Queries to JuruXML may include tagged terms, which will only match similarly tagged instances in the index. We use

$\langle person\rangle$ *and* $\langle role\rangle$

tags in the queries to perform word sense disambiguation of two types: to differentiate a person from, for example, a location with the same name (e.g., Newton - a physicist and Newton - a town in Massachusetts); and to differentiate two different people with the same name belonging to different occupations (e.g., Louis Armstrong a singer and Lance Armstrong an athlete). The second issue can be partially avoided by submitting the full name of a person, but it reduces the amount of documents retrieved about this person dramatically. Besides, as we are interested in creating test data we prefer to retrieve all the documents about people by submitting the query

$\langle person\rangle Name\langle/person\rangle\ \langle role\rangle Occupation\langle/role\rangle$.

The number of documents retrieved varied from one, for the query

$\langle person\rangle Cauchy\langle/person\rangle\ \langle role\rangle mathematician\langle/role\rangle$

to up to 8,144, for

$\langle person\rangle Clinton\langle/person\rangle\ \langle role\rangle politician\langle/role\rangle$.

To counteract misbalance in the data we relied on the $tf * idf$ ranking of JuruXML to sort the matching documents. The top ten such documents were kept (or all of them if fewer than ten were returned).

6.2.2 Extracting Occupation-Specific Activities

To automatically discover general and occupation-specific activities we use a modified version of atomic relations. Atomic relations are triplets consisting of two named entities and a

First Named Entity	Verb	Second Named Entity
Columbus/PERSON	died/VBN	1506/DATE
Columbus/PERSON	sailed/VBD	India/PLACE

Table 6.2: A sample of atomic relations extracted for the collection of documents about Christopher Columbus

verb or an action-defining noun which labels the relation between these two named entities. We extract 189 lists of atomic relations according to the following procedure:

1. For each person analyze the corresponding collection of documents retrieved for this person.

2. From every sentence containing the name of the person under analysis extract all the pairs of named entities, one of the elements of which is the name of this person.

3. For every such pair of named entities extract all verbs, excluding modal and auxiliary verbs, that appear in-between them.

The NE tagger we use is a derivative of that described in (Prager, Chu-Carroll, Brown & Czuba 2006). It tags named entities of about 100 types. Some of the marked types are very specific, like ZIPCODE and ROYALTY. To avoid overfitting we choose six high-level types for atomic relations extraction: PERSON, PLACE, DATE, WHOLENO, ORG and ROLE. We analyze all the atomic relations extracted for different people belonging to the same occupation. A description of each person is a separate collection of documents. Thus, we cannot use the original atomic relation scores described in 3, instead, we keep simple counts for the triplets as later we combine triplets extracted for different people. This technique is similar to the one used for domain template induction where each domain instance description was represented by a separate document collection. Table 6.2 contains examples from the list of atomic relations extracted for Columbus.

First Named Entity	Verb	Second Named Entity
Vespucci/PERSON	explored/VBD	S. America/PLACE
Bering/PERSON	explored/VBD	Aleutian/PLACE

Table 6.3: A sample of atomic relations extracted for two representatives of the explorer occupation

6.2.3 Generalized Atomic Relations

Our goal is to collect information about activities general for all people and about activities specific for some occupations. Thus, we are interested in the semantic information about atomic relations rather than in the relation between the exact named entities encoded by each relation. We made the same assumption for the domain template induction procedure in Chapter 4. Thus, we analyze not the atomic relations themselves but the generalized versions of the extracted atomic relations. The generalized atomic relations identify to what types of named entities people of various occupations are linked through the activities (verb and action relation nouns) they perform. For example, here are two sentences about explorers:

> Vespucci explored the shores of South America.
>
> Vitus Bering explored Aleutian Islands.

The corresponding atomic relations extracted for these sentences are presented in Table 6.3. Clearly, these atomic relations capture information about the same type of activity, namely that explorers explore various locations. Thus, these two atomic relations describe the same activity. What makes these atomic relations different is the exact names of the explorers and the locations a particular explorer explored. We can unify these atomic relations by omitting the exact named entities and leaving only their types. The resulting atomic relations we call *generalized atomic relations*. The atomic relations presented in Table 6.3 can be converged to the following generalized atomic relation:

> NAME/PERSON - explored/VBD - PLACE

In the generalized atomic relations we distinguish two types of named entities with the tag PERSON: those which refer to the person under analysis (from now on they are marked as NAME/PERSON) and all the rest (marked as PERSON). Thus, we separate the person under analysis from the people who are linked to this person through various activities. This generalization technique is similar to the one used by researchers for semantic pattern discovery for information extraction (Collier 1998, Yangarber 2003, Barzilay & Lee 2003, Sudo et al. 2003) and has been successfully applied for the domain template induction procedure described in Chapter 4.

In Chapter 3 we showed that atomic relations capture the most important relations described in a text collection and assign to them good-quality labels. In this work we show that generalized atomic relations can be used for capturing the activities performed by people of different occupations. These activities can be used for describing people's life taking into consideration their occupations.

To select occupation-related activities we merge lists of atomic relations corresponding to the people of the same occupation. Hence, we get ten lists of generalized atomic relations corresponding to the ten occupations under analysis. The count of each generalized atomic relation is equal to the sum of the counts of all the atomic relations which are merged into this generalized atomic relation.

6.3 Getting Occupation-Related Activities

We assume that the activities important for an occupation are linked to the named entities important for this occupation and vice versa, the named entities important for this occupation are linked to the representatives of this occupation through the important activities. Formulated like this, the problem of identifying the actions important for an occupation can be solved using the methodology suggested by Kleinberg (1999) for ranking web-sites, where a search engine counts "inbound and outbound links to identify central sites in a community." The major idea of this technique is based on the assumption that good hubs contain links to good authorities and that links to good authorities are listed within good hubs. Treating activity verbs as hubs and named entity tags as authorities, we map the

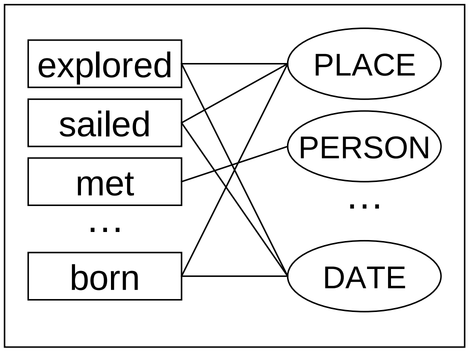


Figure 6.2: Bipartite graph for a set of generalized atomic relations corresponding to *explorers* occupation

problem of discovering activities closely related to a specific occupation to the problem of ranking the reliability of web-pages for the submitted query.

To rank the importance of activities for a particular occupation we define a bipartite graph $G = \{N, V, E\}$, where V are the verb nodes (activities), N are the nodes corresponding to the named entity types linked to V verbs, and E are the arcs connecting the named entity types and the activities. A part of such a bipartite graph created for the *explorers* occupation is presented in Figure 6.2.

Following this procedure, we create bipartite graphs for every occupation. Each of the created ten bipartite graphs contains m named entity types on one side and k verbs (activities) on the other side.[4]

We define $P_{N \to V}$ as a $m \times k$ stochastic transition matrix from named entities to verbs, with elements[5]

$$P_{N \to V}[i,j] \equiv p_{n_i,v_j} = (1-c)\frac{f(n_i \to v_j)}{\sum_{v \in V} f(n_i \to v)} + c \tag{6.1}$$

where f is equal to the sum of the counts of all the generalized atomic relations containing this link for the occupation under analysis. In the same way we define $P_{V \to N}$ $k \times m$ row-stochastic transition matrix from verbs to named entities. Using $P_{V \to N}$ and $P_{N \to V}$, we can define the transition matrix:

[4] Variables m and k are unique for every occupation

[5] To avoid data sparseness we use a smoothing factor $c = 0.01$.

Dancer	Physicist	Singer
made/VBD	born/VBN	said/VBD
died/VBD	died/VBD	born/VBN
appeared/VBD	announced/VBD	died/VBD
been/VBN	discovered/VBD	join/VB
founded/VBD	be/VB	singing/VBG
became/VBD	including/VBG	sang/VBD
born/VBN	became/VBD	has/VBZ
danced/VBD	wrote/VBD	conducting/VBG
blessed/VBN	helped/VBD	made/VBD
perform/VB	named/VBN	became/VBD

Table 6.4: Top ten activities for three occupations: dancers, physicists and singers

$$P_{V \to V} = P_{V \to N} \cdot P_{N \to V} \tag{6.2}$$

that can be used for scoring the verbs according to how important they are for the current occupation. Due to the construction rules, this matrix is a square stochastic matrix where the sum of all the elements in a each row and each column is equal to one.

According to Markov Chain Theory (Kemeny & Snell 1960), for a square stochastic matrix it is possible to find a steady state which corresponds to the eigenvector for the eigenvalue equal to 1. Any square stochastic matrix has 1 among its eigenvalues. The same way the eigenvector corresponding to the steady state for web-pages ranks these pages, the eigenvector corresponding to the steady state of transition matrix (6.2) ranks how tightly the activities are linked to the occupation under consideration. The size of this matrix depends on the variety of the verbs in all forms used in the generalized atomic relations for the representatives of this occupation and varies from 800 for physicists up to 2100 for politicians in our data. This transition matrix models the flow of an imaginary text. For this imaginary text, a description of the activity which should be output next depends on the previously output activity.

Table 6.4 contains top ten activities for three occupations: dancers, physicists and singers. These activities are listed in the sorted order, the ones on top of the table have the highest scores in the respective eigenvectors corresponding to the eigenvalues of 1.

The activities presented in Table 6.4 can be divided into three types:

1. those which are occupation-specific, such as *danced* and *perform* for dancers, *discovered* for physicists, *singing* and *sang* for singers;

2. those which are likely to be used in any biography, such as *born*, *died*, *became*;

3. other, which are mostly general purpose verbs, such as *been*, *made*.

The goal of four classification task is to assign a person to her respective occupation. Thus, for our classification we rely mainly on the first type of activities, i.e. those activities that capture the description of a person as a practitioner of a particular occupation. To extract the activities of the second type we create $P_{V \to V}$, a transition matrix for the combined set for all the generalized atomic relations created for all the ten occupations. According to the matrix construction rules that we describe above, this matrix is also stochastic and by calculating the eigenvector corresponding to its steady state we can identify those activities which are tightly linked to any person irrespectively of his/her occupation and thus reflect general biographical information. Table 6.5 contains top ten activities for this matrix.

In Section 6.4 we show that the lists of occupation-related and general activities are reliable features for classifying people according to their occupations.

6.4 Classification

In this section we describe people classification according to their occupations. For our classification experiments we use a multi-class SVM classifier.[6] As we have 189 data-points corresponding to ten classes we use leave-one-out cross validation which allows us to use the maximal possible amount of data for training. We experiment with two sets of features: one set consists only of the verbs corresponding to the occupation-specific activities (Section 6.4.1); the other set consists of the complete triplets for the generalized atomic relations (Section 6.4.2).

[6] `http://www.cs.cornell.edu/People/tj/svm_light/svm_multiclass.html`

Biography-related verbs
said/VBD
born/VBN
died/VBD
wrote/VBD
became/VBD
had/VBD
known/VBN
be/VB
included/VBD
including/VBG

Table 6.5: Top ten activities common for all the eleven occupations.

6.4.1 SVM Classification Using Only Verbs

To get verb features for multi-class SVM classification we use ten occupation-related lists of activities. Each of these lists is a sorted list of verbs for a respective occupation. The verbs are sorted according to the respective values of the eigenvector corresponding to the steady state of the respective occupation matrix.

The verb-only algorithm is as follows:

V1 Get the sorted list of activities (verbs) for every occupation (ten lists). These activities are the major features on which SVM relies to assign an occupation to a person.

V2 Get the sorted list of activities for all occupations merged together. These activities are used in Step V4 to remove from the list of classification features those activities which are general and not helpful for identifying the occupation of a person.

V3 Get the top 15% of the activities from each of the ten occupation-specific lists and the list of general activities.

V4 From the ten occupation-specific lists remove those activities which are also present in the list of general activities.

			SVM classification				MI		Random
			Verbs		Atomic Relations				
Occupation	# of Reps	Average # of Docs	#	Ratio	#	Ratio	#	Ratio	
Artists	20	10.0	9	0.450	15	0.750	14	0.700	0.106
Athletes	18	10.0	12	0.667	16	0.889	14	0.778	0.095
Composers	20	9.65	10	0.500	15	0.750	19	0.950	0.106
Dancers	15	9.07	7	0.467	13	0.867	11	0.733	0.079
Explorers	17	9.0	12	0.706	15	0.882	15	0.882	0.090
Mathematic.	20	7.2	10	0.500	8	0.381	20	1.000	0.106
Physicists	19	7.05	5	0.263	6	0.316	13	0.684	0.101
Politicians	20	10.0	9	0.450	12	0.600	1	0.050	0.106
Singers	20	9.05	9	0.450	12	0.600	10	0.500	0.106
Writers	20	10.0	7	0.350	12	0.600	10	0.500	0.106
Average				0.480		0.663		0.677	

Table 6.6: Performance of different classification methods.

V5 Merge ten occupation-related lists into one list and remove from this list all the activities that appear in more than 2 occupations.

By leaving at Step 3 some percentage of the activities (verbs) instead of an absolute amount, we take into account the fact that the number of activities used to describe different occupations varies from occupation to occupation (for example, 1794 activities are used in the atomic relations for composers, and 800 - for physicists). As the activities get scores according to the steady state vectors, the activities with high scores are the ones which are most likely to be used for the description of a person of the current occupation. The activities with low values are too specific and are likely to be used in only a few descriptions of people of this occupation. For example, we know that Alexander Borodin was both a composer and a chemist: we do not want to keep those specific verbs which describe his activities as a chemist in the list of the activities describing composers.

In Step 4 we remove from our final list those activities that are typical for all humans and thus cannot be used to distinguish among different occupations. In Step 5 we make our activities as specific as possible: For example, there will be some intersection in activities among mathematicians and physicists, and such activities cannot be helpful for differentiation between these occupations.

The final activities list is used as the list of features for SVM classification. Then we assign values to these features for every person: if the activity from the features list is used as a connector for the extracted atomic relations, then this feature receives the value of 1, if there is no atomic relation using this activity as a connector then this feature is assigned 0. We use binary values for our features instead of the atomic relation counts because the reliability of the scores for the atomic relations extracted for different people varies greatly. For some people we retrieve 10 documents with many biographical facts about those people, for other people we retrieve 2 or 3 documents which only mention the people queried.

Removal of some of the features is reinforced by the (Koller & Sahami 1997) work on feature selection for document classification. It indicates that keeping only a small fraction of the available features improves the classification performance. The optimal number of features is still to be determined.

We train our classifier and evaluate its performance using leave-one-out cross validation. Out of 189 people, eight are not assigned any features. This is because all the atomic relations extracted for these 8 people are either too general or too specific. As these 8 people do not have any features to assign them to the most likely occupation, they are misclassified to the default occupation (artists). Six of these eight are mathematicians, one is a dancer, one is a physicist. Absence of verbal features can be explained by the small number of the documents retrieved for these people. In fact, that less than or equal to three documents were retrieved for these people. Table 6.6 shows how many documents are analyzed per person on average for each occupation. Due to the nature of our document collections, the smallest number of documents analyzed was for mathematicians and physicists: current newswire texts do not contain much information about scientists, and those parts of the World Book and Encyclopedia Britannica which we had at our disposal only contained information for some of the scientists. Out of the remaining 181 people, only 90 are classified correctly. As the distribution of people across occupations is not even, Table 6.6 contains two numbers for each occupation: the absolute number and the ratio of people classified correctly for this occupation.

We believe that the performance of SVM classification based solely on the activities is poor because it does not take into account the information that many activities which are expressed with the help of the same verb are surrounded by different types of arguments for different occupations. For example, Henri Matisse is classified as a dancer based on the frequent co-occurrence with the *dance* activity, which is understandable as one of his most famous paintings is “Dance”. Or, the *explored* activity is among the top activities for several occupations: writers, composers, explorers; but this activity is linked to the PLACE named entity tag only for the explorers. Though the classification based solely on verbs gives quite poor results we consider it to be a valid starting classification as usually activities are associated with the verbs corresponding to these activities.

6.4.2 SVM Classification Using Atomic Relations

To create generalized atomic relation features for multi-class SVM classification we use the sorted lists of activities for the ten occupations and the general list of activities. The activities are sorted according to the values they get from the eigenvector corresponding to the steady state.

AR1 Same as step V1 above.

AR2 Same as step V2 above.

AR3 For the top 15% of the activities from the ten occupation-specific lists get all the generalized atomic relations containing those activities.

AR4 For the top 15% of the activities typical for all the occupation (Step AR2) get all the generalized atomic relations containing those activities.

AR5 From the ten occupation-related lists (Step AR3) remove those generalized atomic relations which are also present in the list of general generalized atomic relations (Step AR4).

AR6 Merge the ten occupation-related lists into one and remove from it all the generalized atomic relations that appear in more than 2 occupations.

Out of 189 people, nine are not assigned any features. This again is because all the atomic relations extracted for these 9 people were either too general or too specific. The people who do not get any relation features are the same as those who do not get any verb features plus one physicist. Out of the remaining 180 people, 124 people are classified into the appropriate occupations. Table 6.6 shows that generalized atomic relations are more reliable for occupation classification than plain activities extracted from these generalized atomic relations. Thus, it can be concluded that structured information captured by atomic relations is valuable and reliable. For example, using atomic relations, Henri Matisse was correctly classified as an artist. According to the t-test the performance of the classification based on atomic relations is significantly better ($p < 0.05$) than the performance of the classification based solely on activities.

We would like to note, that after closer analysis some of the cases of misclassification can be considered as correct assignments as a person could excel in different occupations. For example, in our corpus Paul McCartney is defined as a singer, but classifying him as a composer is a valid result as well.

6.4.3 Other Types of Classification

The task of classifying people according to their occupations is new and to our knowledge there is no existing baseline we could compare our results with. Thus, we decided to adapt for comparison two classification techniques used for other tasks: random assignment of an occupation and mutual information for a name of a person co-occurring with a title an occupation.

Random occupation assignment. As the distribution of people among the occupations is not even we cannot give one exact probability of assigning a correct occupation to a particular person. Instead, we calculate such random probabilities for each occupation. The results are presented in Table 6.6.

Random assignment gives a very low baseline which we easily outperform. For this reason, we use another classification based on mutual information to estimate how good our results are. Classification based on mutual information is considered to be the state-of-the-art classification method for such tasks as hidden web classification (Ipeirotis, Gravano & Sahami 2001) and answer verification (Magnini, Negri, Prevete & Tanev 2002).

Mutual Information (MI). First, we get the counts of how many documents are retrieved for the queries containing only the titles of the occupations (e.g., "⟨*role*⟩ *mathematician* ⟨*/role*⟩", "⟨*role*⟩ *artist* ⟨*/role*⟩", etc.). Then, we get the counts for the queries containing all possible combinations of people's names and occupations' titles. (for example, "⟨*role*⟩ *mathematician* ⟨*/role*⟩ ⟨*person*⟩ *Picasso*⟨*/person*⟩", "⟨*role*⟩ *artist* ⟨*/role*⟩ ⟨*person*⟩ *Picasso* ⟨*/person*⟩", etc.). Finally, we divide the counts for the queries submitted for the occupation plus person by the count for the corresponding occupation query. The maximum of all the ratios for the person gives the occupation for this person.

$$Occupation_j = max_{for\ all\ i,j} \frac{count_{occupation_i,\ person_j}}{count_{occupation_i}} \tag{6.3}$$

According to Table 6.6 SVM, classification based on atomic relations outperforms MI classification for six occupations out of ten, for one occupation (*explorers*) the results for SVM classification and MI are the same and for three occupations MI classification outperforms SVM classification. One of the cases where MI classification outperforms SVM classification is *mathematicians*, where nine mathematicians have no features in SVM classification and thus, do not have any better than random chance to be classified correctly.

Thus, our SVM classification of people according to their occupations based on atomic relations has performance comparable to MI-based classification. This is significant since for many NLP-related applications, MI classification outperforms other methods and is considered to be one of the most powerful classification methods (Ipeirotis et al. 2001, Magnini et al. 2002).

6.4.4 Using Classification Extracted Features

Though we do not dramatically outperform MI, our methodology has one crucial advantage. We use classification not as a primary task but as an evaluation testbed to show that the lists of generalized atomic relations created for every occupation and for general biographies indeed capture the major activities performed by people of the respective occupations and can be used for biography generation. For example, the generalized atomic relations which are used for the description of representatives within all the ten occupations and are excluded from the list of features for SVM classification as too general, contain verbs such as *born/VBN*, *died/VBD* linked to the DATE and PLACE named entity tags or *became/VBD* linked to the ROLE named entity tag. Table 6.7, on the other hand, contains occupation-specific generalized atomic relations. These generalized atomic relations have high scores within the respective occupations, are used as features for SVM classification and have non-zero values in the feature sets which correctly classified people into the appropriate occupations.

Table 6.8 contains a sample of generalized atomic relations which were used for the representatives of all the ten occupations analyzed and thus, can be used for either identifying the snippets of text which contain biographical information or they can be used for constructing auxiliary questions.

In this chapter, we have shown, that the occupation-related activities learned from descriptions of sets of people belonging to a pre-defined set of occupations can be successfully used as classification features to assign a new person to her respective occupation from the initial occupation list. Moreover, if people under analysis belong to a variety of occupations, using the same methodology we can learn general biographical activities that are likely to be used in a description of any person irrespectively of her occupation. One can notice, that we can map the task of activities learning to the task of domain modelling (Chapter 4). Thus, the activities that we learned in unsupervised fashion as described above, can be used to create domain templates corresponding either to the occupation under analysis or to a generic biography.

In the rest of the chapter, we explore the possibility of connecting domain templates (sets of occupation-related activities) among each other. We describe clustering procedures within different input number of clusters and show, that the activities extracted from people descriptions can be used for grouping people according to their occupations. Afterwards, we show how clustering can be used to create hierarchies of domains.

6.5 Clustering

By definition, a classification task requires a pre-defined set of the classes into which the input elements should be classified. In most cases, including human occupations, it is not possible to come up with an exhaustive list of such classes. Thus, many tasks use clustering

Artists	Athletes
NAME - painted/VBN - DATE	NAME - win/VB - WHOLENO
NAME - resemble/VB - PERSON	NAME - scored/VBD - WHOLENO
PERSON - designed/VBN - NAME	NAME - winning/VBG - DATE
Composers	Dancers
NAME - composed/VBN - PERSON	NAME - danced/VBN - ORG
NAME - include/VBP - WHOLENO	PLACE - presented/VBD - NAME
ROLE - hearing/VBG - NAME	NAME - appeared/VBD - WHOLENO
Explorers	Mathematicians
NAME - annexes/VBZ - PLACE	PERSON - developed/VBD - NAME
NAME - reach/VB - PLACE	ROLE - prove/VB - NAME
NAME - declares/VBZ - DATE	NAME - studied/VBD - WHOLENO
Physicists	Politicians
DATE - described/VBD - NAME	NAME assassinated/VBN - PLACE
ROLE - predicted/VBD - NAME	NAME - postponed/VBN - PLACE
NAME - continued/VBD - ORG	NAME - flown/VBN - ORG
Singers	Writers
NAME - conducting/VBG - PLACE	PLACE - leaving/VBG - NAME
NAME - sing/VB - ROLE	NAME - translated/VBN - PERSON
NAME - sang/VBD - PERSON	WHOLENO - written/VBN - NAME

Table 6.7: Occupation-specific generalized atomic relations (note that name stands for name/person).

First Named Entity	Verb	Second Named Entity
NAME/PERSON	born/VBN	PLACE
NAME/PERSON	died/VBD	DATE
NAME/PERSON	became/VBD	ROLE

Table 6.8: Generalize atomic events typical for any biography

rather than classification to group closely related elements. In this section, we describe how clustering can be used for grouping people according to the activities used in the descriptions of these people. We also show that clustering can be used not only for identifying activities typical for some particular occupations but also for identifying which activities can be used to describe people of several occupations and, propagating further, which activities are general biographical ones. Thus, in contrast to the classification experiment, we do not eliminate general biographical activities from the features list. Rather, we use ***all generalized activities*** identified in the people's description and show that general biographical and occupation-specific activities can be identified automatically as a side-effect of the clustering process.

The generated hierarchy of clusters links those occupations that have intersecting sets of occupation-specific activities. This hierarchy contains information about various human occupations together with practitioners of each occupation and a set of activities which are used in the descriptions of people belonging to each particular occupation. The root node of this hierarchy contains activities which verbalize general biographical information and can be used in a biography of any person. Thus, the final hierarchy, which is learned in an unsupervised fashion from unstructured data (text), can be used as an initial step for building an ontology of human occupations. At the same time, the described methodology can be potentially used to tease apart descriptions of people having the same name but belonging to different occupations (Artiles, Gonzalo & Verdejo 2005).

6.5.1 CLUTO CLustering Tool

For our experiments, we used the CLUTO clustering toolkit which has been shown to produce high-quality clustering solutions (Zhao & Karypis 2001). In our experiments, we used k -way clustering with a *direct* clustering procedure when all k clusters are formed simultaneously. We use the following clustering criterion function:

$$\mathcal{I}_2 = \sum_{i=1}^{k} \sqrt{\sum_{v,u \in S_t} sim(u,v)}$$

where k is the total number of clusters, S is the total number of objects to be clustered, u and v represent two objects from S, S_i is the set of objects assigned to the ith cluster, $sim(u,v)$

is the similarity function between two objects (in our case *cosine* similarity between two vectors corresponding to two people).

We used the activities extracted in Section 6.2.1 as features for the clustering experiments. This data was encoded in a $N \times M$ matrix, where N was the number of people to be clustered (189) and M was the number of unique activities (generalized atomic events) used in the description of all the 189 people (N is equal to 21433). The elements of the $N \times M$ matrix are assigned 0 or 1 according to the following criterion:

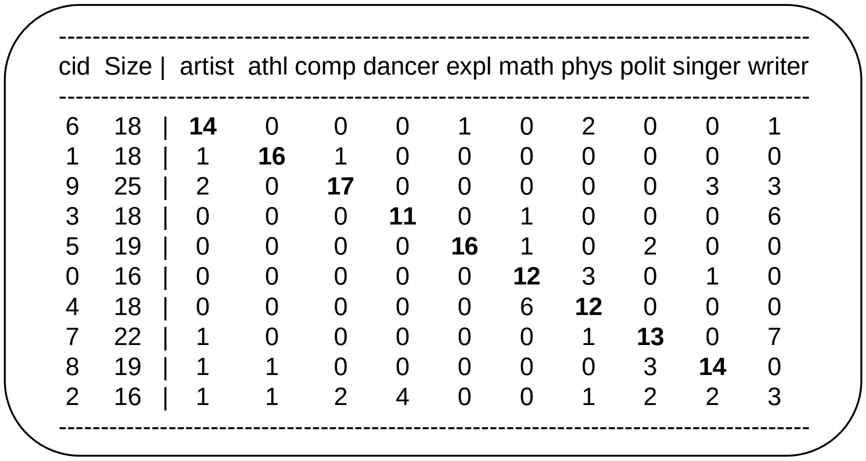
$$a_{i,j} = \begin{cases} 1, & \text{if the activity } j \text{ is used in the} \\ & \text{description of the person } i; \\ 0, & \text{otherwise.} \end{cases}$$

We used binary values for our features instead of statistics about the activities for each person because the document collections retrieved differed in sizes and quality.

6.5.2 Identifying the Number of Clusters

Any clustering algorithm assumes the knowledge of a pre-defined parameter which either defines how close the elements of the clusters are or the number of clusters that should be created. We run several experiments grouping people into different number of clusters. We show, that when the number of clusters is chosen to be equal to the number of the underlying occupations, then clustering can be used for automatic grouping of people belonging to the same occupation. At the same time, we show how the result of clustering, where the number of outcome clusters is smaller than the number of the underlying occupations, can be used for grouping occupations having similar activities. Thus, we demonstrate a possibility of using clustering for creating domain hierarchies. We also compare the grouping of people inside clusters against the gold standard occupation assignment.

Ten Clusters. Identifying the optimal number of clusters into which the input data should be broken is a difficult task. In our corpus there are representatives of ten occupations. Some of these occupations are very distinct from the rest of the occupations and are not likely to have a lot of occupation-related activities used by people outside of these occupations (e.g. dancers, explorers). On the other hand, some occupations are quite close and can



```
----------------------------------------------------------------------------------
cid  Size |  artist  athl comp dancer expl math phys polit singer writer
----------------------------------------------------------------------------------
6    18  |  14     0     0     0     1     0     2     0     0     1
1    18  |  1     16     1     0     0     0     0     0     0     0
9    25  |  2      0    17     0     0     0     0     0     3     3
3    18  |  0      0     0    11     0     1     0     0     0     6
5    19  |  0      0     0     0    16     1     0     2     0     0
0    16  |  0      0     0     0     0    12     3     0     1     0
4    18  |  0      0     0     0     0     6    12     0     0     0
7    22  |  1      0     0     0     0     0     1    13     0     7
8    19  |  1      1     0     0     0     0     0     3    14     0
2    16  |  1      1     2     4     0     0     1     2     2     3
----------------------------------------------------------------------------------
```

Figure 6.3: Grouping people into ten clusters.

have substantial intersections of the lists of occupation-related activities (e.g., physicists and mathematicians). This issue complicates the task of choosing the optimal number of clusters.

For our initial experiment, we assume that the optimal number of clusters corresponds to the number of occupations covered in our corpus. We also assume that we know in advance how many occupations are covered in our corpus. Thus, for the initial clustering experiment, the input number of clusters into which people should be divided is equal to ten.

Figure 6.3 contains the result of running the clustering algorithm with *ten* as a parameter identifying the number of clusters into which the input elements (people) should be divided. In Figure 6.3 we ordered the clusters in such a way so that the diagonal is maximized (i.e., the row where ith element is maximal is placed in the ith row). The identification numbers for clusters presented in the *cid* column are the same as they are produced by the CLUTO clustering tool, where "clusters that are *tight* and far away from the rest of the objects have smaller cid values." This manual re-ordering of clusters is performed solely for the visualization purpose and it does not modify any of the numbers computed by the clustering procedure. The tightness of clusters is calculated as difference between internal and external similarities, where internal similarity is the average similarity between the objects in the cluster and external similarity is the average similarity of the objects of some cluster and the rest of the objects from the other clusters.

One can see that all the created clusters except one (Cluster 2) have a dominant occupation which can be used as a label for the respective cluster. For example, 14 out of 18 members of Cluster 6 are artists. For Cluster 2, whose elements are spread among several occupations, it is not possible to pinpoint one dominant occupation.

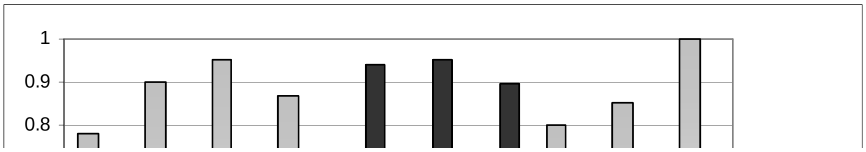
The same can be noticed about occupations. For most occupations it is easy to point out the cluster containing the vast majority of its representatives. For example, 16 out of 17 explorers are placed into Cluster 5. Only writers are spread among five clusters, with no cluster distinctively standing out as having the vast majority of writers.

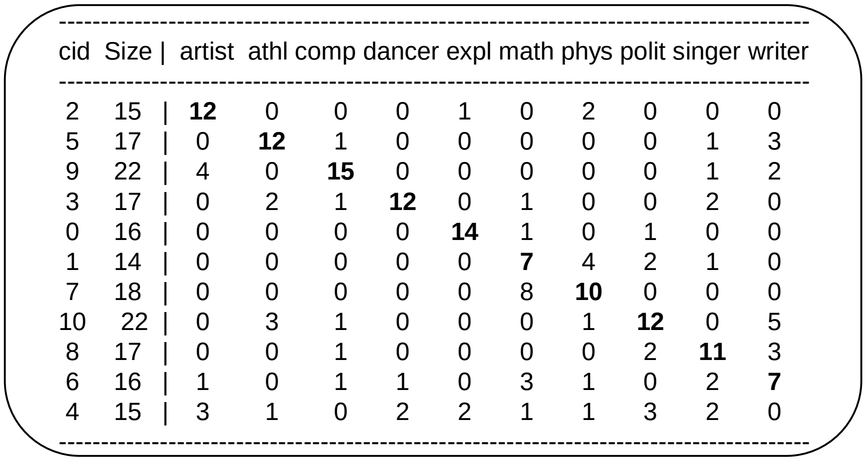
Using CLUTO it is possible to analyze the distribution of clustering features, and to investigate which activities are the most descriptive and most discriminative ones. In our case, we are interested in analyzing which activities are related to which clusters. Table 6.9 contains sets of descriptive features (activities) for each cluster. These activities are maximal cliques for the graph, where the features (activities) "are connected via an edge if and only if their co-occurrence frequency within the cluster is greater than their expected co-occurrence" (Zhao & Karypis 2001).

Two Clusters. It must be noted, that the problem of estimating the optimal number of clusters for a data set is one of the most essential issues in cluster analysis. We do not tackle this problem in this work. Rather, we show that for the problem of clustering people according to their occupations even a non-optimal number of clusters gives meaningful and useful results.

Figure 6.4 shows the distribution of people between two clusters. In Figure 6.4 the separation between explorers, mathematicians and physicists on the one side, and representatives of other occupations on the other side is clear-cut. This distribution can be roughly mapped to the division of the occupations used in our experiments into *Arts* and *Sciences*.

Eleven Clusters. At the same time, the result of the clustering procedure with the number of clusters greater than ten still gives meaningful results. For example, the result of grouping people into eleven clusters is presented in Figure 6.5. One can see that physicists and mathematicians are evenly spread between two Clusters 1 and 7, which emphasizes the closeness of these two occupations.





cid	Size	artist	athl	comp	dancer	expl	math	phys	polit	singer	writer
2	15	**12**	0	0	0	1	0	2	0	0	0
5	17	0	**12**	1	0	0	0	0	0	1	3
9	22	4	0	**15**	0	0	0	0	0	1	2
3	17	0	2	1	**12**	0	1	0	0	2	0
0	16	0	0	0	0	**14**	1	0	1	0	0
1	14	0	0	0	0	0	**7**	4	2	1	0
7	18	0	0	0	0	0	8	**10**	0	0	0
10	22	0	3	1	0	0	0	1	**12**	0	5
8	17	0	0	1	0	0	0	0	2	**11**	3
6	16	1	0	1	1	0	3	1	0	2	**7**
4	15	3	1	0	2	2	1	1	3	2	0

Figure 6.5: Grouping people into eleven clusters.

Thus, if we choose the number of clusters to be less than the actual number of occupations, we group together those occupations which are closely related. On the other hand, choosing the number of clusters greater than the actual number of occupations can cause breaking occupation clusters even further. We speculate that it is possible to capture the point where increasing the number of clusters is erroneous. For example, one can see that in Figure 6.5, the number of clusters without a distinct row dominant is three (clusters with *cid*s 1, 4, and 7). While in Figure 6.3, there is only one cluster without a distinct dominant (cluster with *cid* 2). The problem of getting the optimal number of hierarchy levels, however, is outside of the scope of this thesis.

6.6 Conclusions and Future Work

In this chapter, we reported results on how atomic relations can be used for classifying people according to their occupations. We introduced a novel representation for describing human activities (generalized atomic relations). We showed how random walk theory can be used for identifying general biographical and occupation-specific activities. We described experiments using SVM classification together with generalized occupation-specific atomic relations as features and showed, that the results of this classification were comparable to the results of state-of-the-art classification techniques.

We also showed that generalized atomic relations can be used not only for classification but for clustering as well. In our clustering experiment we introduced a second approach for unsupervised learning of general biographical and occupation-specific activities. We proposed a way of creating a hierarchy of occupations where each node represents a set of closely related occupations together with the list of representatives of these occupations and occupation-related activities shared by the node occupations. The leaf nodes of this hierarchy describe unique occupations, and the root has a list of general biographical activities.

In the future, we are interested in investigating ways of identifying other types of activities which are neither general no occupation-specific but rather person-specific. We also plan to put into hierarchies other domains and activities corresponding to these domains. For example, disasters can be divided into natural disasters and technological disasters, while natural disasters can be divided into wind-related (tornados, hurricanes, etc.), water-related (tsunamis, floods, etc.), and other (earthquakes, volcano eruptions, etc.), and so on. We also believe that the usage of generalized atomic relations can enable significant new techniques for a number of natural language processing tasks (i.e., QA, biographical summarization, IR query expansion, etc.).

cid		Activities
6	22.22%	NAME-called/VBD-/ORG; NAME-running/VBG-/DATE; NAME-led/VBD-/PLACE
	30.56%	NAME-painted/VBN-/DATE; NAME-called/VBD-/ORG
	25.00%	NAME-sold/VBN-/ROLE; NAME-running/VBG-/DATE
1	27.78%	NAME-win/VB-/WHOLENO; NAME-skating/VBG-/ROLE /ROLE-skating/VBG-NAME ;/PERSON-evert/VBP-NAME
	42.59%	NAME-win/VB-/WHOLENO; NAME-won/VBD-/WHOLENO; NAME-skating/VBG-/ROLE
9	55.00%	NAME-composed/VBN-/DATE; NAME-written/VBN-/ROLE NAME-born/VBN-/ROLE; NAME-wrote/VBD-/WHOLENO
	48.00%	NAME-composed/VBN-/PERSON; NAME-composed/VBN-/DATE; NAME-written/VBN-/ROLE; NAME-wrote/VBD-/WHOLENO
3	22.22%	/PLACE-dancing/VBG-NAME
	33.33%	NAME-dance/VB-/PERSON; NAME-appeared/VBD-/PLACE
	16.67%	NAME-introduced/VBN-/PLACE; NAME-introduced/VBN-NAME
5	34.21%	NAME-sailed/VBD-/PLACE; NAME-sailed/VBD-/DATE; NAME-explore/VB-/PLACE; NAME-explored/VBN-/PLACE
	42.11%	NAME-sailed/VBD-/PLACE; NAME-discover/VB-/PLACE; NAME-sailed/VBD-/DATE
0	12.50%	/DATE-named/VBD-NAME; NAME-stated/VBD-/WHOLENO
	9.38%	/ROLE-discovered/VBN-NAME; /ROLE-turns/VBZ-NAME
	15.62%	/ROLE-discovered/VBN-NAME; /DATE-named/VBD-NAME
4	25.93%	NAME-became/VBD-/ORG; /DATE-named/VBN-NAME NAME-named/VBN-/ROLE
	18.06%	NAME-became/VBD-/ORG; NAME-summarized/VBN-/DATE; NAME-states/VBZ-/ROLE; NAME-named/VBN-/ROLE
7	56.06%	NAME-been/VBN-/ORG; NAME-said/VBD-/PLACE; NAME-meet/VB-/PERSON
	63.64%	NAME-been/VBN-/ORG; NAME-said/VBD-/PLACE; /PERSON-told/VBD-NAME
	56.06%	NAME-meet/VB-/PLACE; NAME-said/VBD-/PLACE; NAME-meet/VB-/PERSON
8	25.00%	/PERSON-sang/VBD-NAME; NAME-wasn't/VBD-/ROLE; NAME-sang/VBD-/PLACE NAME-singing/VBG-/PERSON
	26.32%	NAME-perform/VB-/PLACE; /PERSON-sang/VBD-NAME; NAME-sang/VBD-/PLACE; NAME-singing/VBG-/PERSON
2	15.62%	/PLACE-produced/VBN-NAME; NAME-demonstrate/VB-/PLACE
	9.38%	NAME-hides/VBZ-/ORG; NAME-demonstrate/VB-/PLACE
	12.50%	NAME-set/VBN-/PERSON; NAME-played/VBN-/PERSON

Table 6.9: Most descriptive generalized atomic relations (biographical activities) for the clusters presented in Figure 6.3.

Chapter 7

Contributions, Limitations and Future Work

In this thesis, we investigate the problem of identifying, within a text, relations that capture information important for event-focused document collections. The presented solutions work with events of various granularity and show how to use these relations to improve the performance of a number of natural language processing applications. Now, we summarize our main contributions, describe limitations of the presented techniques and delineate directions for interesting future research.

7.1 Contributions

First, we analyzed the concept of *atomic relations*. These are relations between entities that are present in a collection of documents. Combined within a shallow semantic network atomic relations capture the structure of the event described in the input document collection. We showed that shallow semantic network relations, where the central elements of the relations were verbs and action nouns, could be used successfully as information features to capture succinctly the essence of the *events* described in a collection of text documents.

Another important contribution of this dissertation for the summarization task is a formal model of efficient information packaging where the main goal is to cover maximal amount of information within limited space. This model is based on mapping the information selection task onto the set cover complexity problem. Based on this mapping, we provided information selection algorithms that were provably optimal polynomial-time approximations for information selection tasks that have a limit on the output size.

In our work, we further generalized the approach of studying relations among elements within a document collection. We studied sets of document collections describing similar events. We designed and implemented a procedure to infer *domain templates* that could characterize and summarize the important events for a particular *type of event* or *domain*. In effect, the domain templates capture the commonalities across a set of similar events (e.g., earthquakes) and outline the most important aspects of this type of event (domain).

We also studied the problem of creating answers to event-related open-ended questions. We devised an approach that used two different strategies for answer selection depending on the nature of the event in question. When there were enough commonalities of the event in question and a pre-defined domain template, we used event generic properties captured by the domain template to guide the answer selection procedure. However, when the description of the event in question was complex or referred to a unique event, then it was challenging to identify the commonalities between the event in question and any of the pre-defined domain templates; in such cases, our answer selection procedure was guided by a shallow semantic network, generated according to the document sets retrieved for each particular event question.

Another contribution of this thesis is an empirical study indicating that a simple combination of information retrieval and general summarization techniques is not enough for solving an open-ended question-answering problem. We described two error sources for such combination systems, and suggested and evaluated two methods for combining information retrieval and general summarization techniques to work better for question-answering purposes.

Finally, using as an example the biography domain, we devised a procedure for identifying commonalities across different subdomains (in our example, across activities that could be used for descriptions of people belonging to different occupations). As an important contribution, we showed how to use random walk theory to identify biographical information that corresponds to three levels of activity. Our unsupervised learning technique identified automatically three levels of activities: general biographic, occupation-related and person-specific. By building on these ideas, we also showed how to generate hierarchies of human activities with respect to their occupations.

Thus, the main theme of this dissertation is the design and implementation of algorithms for unsupervised learning of important relations for event-focused document collections of various granularity. The proposed algorithms operate over document collections describing various events and use the extracted relations for answer selection for open-ended event-related question-answering and domain modelling tasks. Specifically:

- we designed and implemented the atomic relation extraction algorithm for a document collection;
- we designed and implemented the domain modelling algorithm that identifies regularities in the use of relations across a set of document collections describing similar events (domain instances);
- we outlined a prototype procedure that can be used to identify relations among domain templates and can be used to build a domain hierarchy; we evaluated the results for a proof-of-concept experiment for a domain divided into several subdomains.

7.2 Limitations and Future Work

In this dissertation we showed that unsupervised learning of relations that captured co-occurrences of various elements within a document collection, set of documents collections and domain templates was a powerful tool that could be efficiently used for selecting pieces of text that contained information potentially important for a user. Currently, the relations that we extract do not contain rich semantic and syntactic information. In contrast, manually created MUC templates (Marsh & Perzanowski 1997), case frames (Riloff & Schmelzenbach 1998), PALKA's frame-phrasal pattern structures (Kim & Moldovan 1995), or PropBank annotation schemas (Palmer et al. 2005) contain both syntactic and semantic information. The ideal situation would be to bring together semantic, syntactic and lexical information. We believe that the on-going effort on creating the OntoNotes corpus (Hovy et al. 2006) can provide us with training data for developing tools for combining semantic, syntactic and lexical information for particular text elements. The ability to assign meaningful semantic tags to atomic relations or domain template slots will propel these relations beyond pure co-occurrence estimation. It will also allow us to evaluate the extracted relations and domain template slots manually and effectively eliminate from further consideration those relations that captured noise instead of meaningful information types.

Knowing the semantics of relations can also be used to link the task of relation learning to the task of classic information extraction. On the one hand, those relations that encode the same semantic information can be brought together; on the other hand, these unified relations can be used by IE pattern generation algorithms as initial patterns of a particular semantic type.

Another important issue that is not addressed in this work is identifying on-the-fly those domains for which domain templates should be created. Currently, we use a pre-defined set of domains; in the future, we plan to use the information encoded in the event-related questions.

An interesting research direction with the open-ended event-related QA task is to devise and run a full-fledged user study on what information is expected to be included in the answer and what format of an answer is most helpful for a user. For example, an answer created by the current version of our system is a set of sentences. A user study can be used to investigate the importance of creating a cohesive text out of these sentences. The question to be answered is: if an answer created by a QA system is used by an analyst to create a report on an event, what format of an answer is more helpful: a coherent text or a list of sentences with, perhaps, extended context information. Another goal of such a study is to identify what type of information is most useful for a user and is expected to be present in the answer.

In this dissertation we outlined a prototype procedure that can be used to create a hierarchy of occupations. In the future, we plan to generalize this methodology. We are interested in investigating domain hierarchies outside of the biography domain. For example, disasters can be divided into natural disasters and technological disasters, while natural disasters can be divided into wind-related (tornados, hurricanes, etc.), water-related (tsunamis, floods, etc.), and other (earthquakes, volcano eruptions, etc.), and so on.

Bibliography

Abe, K., Kawasoe, S., Asai, T., Arimura, H. & Arikawa, S. (2002), Optimized substructure discovery for semi-structured data, *in* ‘Proceedings of the 6th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD–2002)’, Helsinki, Finland, pp. 1–14.

Agichtein, E., Cucerzan, S. & Brill, E. (2005), Analysis of factoid questions for effective relation extraction, *in* ‘Proceedings of the 28th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR–2005)’, Salvador, Brazil, pp. 567–568.

Agichtein, E. & Gravano, L. (2000), Snowball: Extracting relations from large plain-text collections, *in* ‘Proceedings of the Fifth ACM International Conference on Digital Libraries’, San Antonio, Texas, USA, pp. 85–94.

Allan, J., Carbonell, J., Doddington, G., Yamron, J. & Yang, Y. (1998), Topic detection and tracking plot study: Final report, *in* ‘Proceedings of the DARPA Broadcast News Trascription Workshop’.

Allan, J., Gupta, R. & Khandelwal, V. (2001), Topic models for summarizing novelty, *in* ‘Proceedins of the Workshop on Language Modeling and Information Retrieval’, Pittsburgh, Pennsylvania, USA.

Androutsopoulos, I., Ritchie, G. & Thanisch, P. (1995), ‘Natural language interfaces to databases– an introduction’, *Journal of Language Engineering* **1**(1), 29–81.

Artiles, J., Gonzalo, J. & Verdejo, F. (2005), A testbed for people searching strategies in the WWW, *in* 'Proceedings of the 28th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR-2005)', Salvador, Brazil, pp. 569–570.

Bach, E. (1986), 'The algebra of events', *Linguistics and Philosophy* **9**, 5–16.

Bagga, A. (1997), Analyzing the complexity of a domain with respect to an Information Extraction task, *in* 'Proceedings of the Seventh MUC Conference'.

Banko, M., Cafarella, M., Soderland, S., Broadhead, M. & Etzioni, O. (2007), Open information extraction from the web, *in* 'Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI–2007)', Hyderabad, India, pp. 2670–2676.

Barzilay, R. & Elhadad, M. (1997), Using lexical chains for text summarization, *in* 'Proceedings of the ACL/EACL–1997 Workshop on Intelligent Scalable Text Summarization', Madrid, Spain, pp. 10–17.

Barzilay, R. & Lee, L. (2003), Learning to paraphrase: An unsupervised approach using multiple-sequence alignment, *in* 'Proceedings of the Joint Human Language Technology Conference and North American Chapter of the Association for Computational Linguistics Meeting (HLT/NAACL–2003)', Edmonton, Canada, pp. 16–23.

Barzilay, R. & McKeown, K. (2005), 'Sentence fusion for multidocument news summarization', *Computational Linguistics* **31**(3), 297–328.

Baxedale, P. (1958), 'Machine-made index for technical literature - an experiment', *IBM Journal of Research and Development* **2**, 354–361.

Berland, M. & Charniak, E. (1999), Finding parts in very large corpora, *in* 'Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL–1999)', College Park, Maryland, USA, pp. 57–64.

Bethard, S. & Martin, J. (2006), Identification of event mentions and their semantic class, *in* 'Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP–2006)', Association for Computational Linguistics, Sydney, Australia, pp. 146–154.

Bikel, D., Schwartz, R. & Weischedel, R. (1999), 'An algorithm that learns what's in a name', *Machine Learning Journal Special Issue on Natural Language Learning* **34**, 211–231.

Biryukov, M., Angheluta, R. & Moens, M.-F. (2005), Multidocument question answering text summarization using topic signatures, *in* 'Proceedings of the 5th Dutch-Belgium Information Retrieval Workshop (DIR–5)', Utrecht, the Netherlands.

Blair-Goldensohn, S., Evans, D., Hatzivassiloglou, V., McKeown, K., Nenkova, A., Passonneau, R., Schiffman, B., Schlaikjer, A., Siddharthan, A. & Siegelman, S. (2004), Columbia University at DUC 2004, *in* 'Proceedings of the 4th Document Understanding Conference (DUC–2004)', Boston, Massachusetts, USA.

Blair-Goldensohn, S., McKeown, K. & Schlaikjer, A. (2003), A hybrid approach for QA track definitional questions, *in* 'the 12th Text Retrieval Conference (TREC–2003)', Gaithersburg, Maryland, USA.

Blair-Goldensohn, S., McKeown, K. & Schlaikjer, A. (2004), Answering definitional questions: A hybrid approach, *in* M. Maybury, ed., 'New Directions In Question Answering', AAAI Press, chapter 4.

Brandow, R., Mitze, K., & Rau, L. F. (1995), 'Automatic condensation of electronic publications by sentence selection', *Information Processing and Management* **31**(5), 675–685.

Brill, E., Dumais, S. & Banko, M. (2002), An analysis of the AskMSR question-answering system, *in* 'Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP–2002)', Philadelphia, Pennsylvania, USA, pp. 257–264.

Brin, S. (1998), Extracting patterns and relations from the World-Wide Web, *in* 'Proceedings of the First International Workshop on the Web and Databases (WebDB–1998)', Valencia, Spain, pp. 172–183.

Bronnenberg, W., Bunt, H., Landsbergen, S., Scha, R., Schoenmakers, W. & Utteren, E. V. (1979), The question answering system PHLIQA1, *in* 'Natural Communication with Computers', Vol. II, Carl Hanser Verlag.

Cafarella, M., Downey, D., Soderland, S. & Etzioni, O. (2005), KnowItNow: Fast, scalable information extraction from the web, *in* 'Proceedings of the Joint Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP–2005)', Vancouver, Canada, pp. 563–570.

Califf, M. & Mooney, R. (1998), Relational learning of pattern-match rules for information extraction, *in* 'Working Notes of AAAI Spring Symposium on Applying Machine Learning to Discourse Processing', Menlo Park, California, USA, pp. 6–11.

Cardie, C. (1997), 'Empirical methods in information extraction', *AI Magazine* **18**(4), 65–79.

Cardie, C., Ng, V., Pierce, D. & Buckley, C. (2000), Examining the role of statistical and linguistic knowledge sources in a general-knowledge question answering system, *in* 'Proceedings of the 6th Applied Natural Language Processing Conference (ANLP–2000)', Seattle, Washington, USA, pp. 180–187.

Carmel, D., Amitay, E., Hersovici, M., Maarek, Y., Petruschka, Y. & Soffer, A. (2001), Juru at TREC-10. Experiments with index pruning, *in* 'Proceedings of the Tenth Text REtrieval Conference (TREC–2001)', Gaithersburg, Maryland, USA.

Chu-Carroll, J., Prager, J., Welty, C., Czuba, K. & Ferrucci, D. (2003), A multi-strategy and multi-source approach to question answering, *in* 'Proceedings of the Twelfth Text REtrieval Conference (TREC–2003)', Gaithersburg, Maryland, USA.

Chung, S. & Timberlake, A. (1985), Tense, aspect, and mood, *in* T. Shopen, ed., 'Language Typology and Syntactic Description', Vol. 3, Cambridge University Press, chapter 4, pp. 202–258.

Clarke, C. L., Cormack, G. V., Lynam, T. R., Li, C. M. & McLearn, G. L. (2001), Web reinforced question answering (MultiText experiments for TREC-2001), *in* 'Proceedings of the Tenth Text REtrieval Conference (TREC–2001)', Gaithersburg, Maryland, USA.

Collier, R. (1998), Automatic Template Creation for Information Extraction, PhD thesis, University of Sheffield, Department of Computer Science.

Daniel, N., Radev, D. & Allison, T. (2003), Sub-event based multi-document summarization, *in* 'Proceedings of the Joint Human Language Technology Conference and North American Chapter of the Association for Computational Linguistics Meeting (HLT/NAACL–2003)', Association for Computational Linguistics, Morristown, New Jersey, USA, pp. 9–16.

Davidson, D., ed. (2001), *Essays in Actions and Events*, Clarendon Press, Oxford, UK.

Doddington, G., Mitchell, A., Przybocki, M., Ramshaw, L., Strassel, S. & Weischedel, R. (2004), The automatic content extraction (ACE) program - tasks, data, and evaluation, *in* 'Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC–2004)', Canary Islands, Spain.

Dolnicar, S. & Grün, B. (2007), 'How constrained a response: A comparison of binary, ordinal and metric answer formats', *Retailing and Consumer Service* **14**(2), 108–122.

Duboue, P. & McKeown, K. (2003), Statistical acquisition of content selection rules for natural language generation, *in* 'Proceedings of the Conference on Empirical Methods for Natural Language Processing (EMNLP–2003)', Sapporo, Japan, pp. 121–128.

Echihabi, A., Hermjakob, U., Hovy, E., Marcu, D., Melz, E. & Ravichandran, D. (2003), Multiple-engine question answering in TextMap, *in* 'Proceedings of the Twelfth Text REtrieval Conference (TREC–2003)', Gaithersburg, Maryland, USA.

Echihabi, A. & Marcu, D. (2003), A noisy-channel approach to question answering, *in* 'Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL–2003)', Sapporo, Japan, pp. 16–23.

Edmundson, H. (1968), 'New methods in automatic extracting', *Journal of the Association for Computing Machinery* **23(1)**, 264–285.

Etzioni, O., Cafarella, M., Downey, D., Kok, S., Popescu, A.-M., Shaked, T., Soderland, S., Weld, D. S. & Yates, A. (2004), Web-scale information extraction in KnowItAll (Preliminary results), *in* 'Proceedings of the 13th International World Wide Web Conference (WWW–2004)', New York, New York, USA, pp. 100–110.

Fellbaum, C., ed. (1998), *WordNet: An Electronic Lexical Database*, MIT press.

Filatova, E. & Hatzivassiloglou, V. (2003), Domain-independent detection, extraction, and labeling of atomic events, *in* 'Proceedings of the Conference on Recent Advances in Natural Language Processing (RANLP–2003)', Borovets, Bulgaria, pp. 145–152.

Filatova, E. & Hovy, E. (2001), Assigning time-stamps to event-clauses, *in* 'Proceedings of the ACL-2001 Workshop on Temporal and Spatial Information Processing', Toulouse, France, pp. 1–8.

Filatova, E. & Prager, J. (2005), Tell me what you do and I'll tell you what you are: Learning occupation-related activities for biographies, *in* 'Proceedings of the EMNLP/HLT Conference', Vancouver, Canada.

Fillmore, C. & Baker, C. (2001), Frame semantics for text understanding, *in* 'Proceedings of the NAACL–2001 Workshop on WordNet and Other Lexical Resources', Pittsburgh, Pennsylvania, USA.

Fiscus, J., Doddington, G., Garofolo, J. & Martin, A. (1999), NIST's 1998 topic detectoin and tracking evaluation (TDT-2), *in* 'Proceedings of the 1999 DARPA Broadcast News Workshop', Herndon, Virginia, USA, pp. 19–24.

Fleischman, M., Echihabi, A. & Hovy, E. (2003), Offline strategies for online question answering: Answering questions before they are asked, *in* 'Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL–2003)', Sapporo, Japan, pp. 1–7.

Gerner, D., Schrodt, P., Francisco, R. & Weddle, J. (1994), 'Machine coding of event data using regional and international sources', *International Studies Quarterly* **38**, 91–119.

Gildea, D. & Jurafsky, D. (2002), 'Automatic labeling of semantic roles', *Computational Linguistics* **28**(3), 245–288.

Goldstein, J., Mittal, V., Carbonell, J. & Callan, J. (2000), Creating and evaluating multi-document sentence extract summaries, *in* 'Proceedings of the Ninth International Conference on Information and Knowledge Management (CIKM–2000)', McLean, Virginia, USA, pp. 165–172.

Green, B. F., Wolf, A. K. & amd Kenneth Laughery, C. C. (1963), Baseall: an automatic question answerer, *in* E. A. Feigenbaum & J. Feldman, eds, 'Computers and Thought', McGraw-Hill, New York, pp. 207–216.

Greenwood, M. A. & Saggion, H. (2004), A pattern based approach to answering factoid, list and definition questions, *in* 'Proceedings of the 7th RIAO Conference (RIAO–2004)', Avignon, France, pp. 232–243.

Grishman, R. (1997), Information extraction: Techniques and challenges, *in* M. T. Pazienza, ed., 'Proceedings of the Information Extraction International Summer School (SCIE–97)', Springer-Verlag.

Grishman, R. (2003), Information extraction, *in* R. Mitkov, ed., 'The Oxford Handbook of Computational Linguistics', Oxford University Press, pp. 545–559.

Guha, R., McCool, R. & Miller, E. (2003), Semantic search, *in* 'Proceedings of the 12th International World Wide Web Conference (WWW–2003)', Budapest, Hungary, pp. 700–709.

Hakkani-Tür, D., Tur, G. & Levit, M. (2007), Exploiting information extraction annotations for document retrieval in distillation tasks, *in* 'Proceedings of the Interspeech 2007 Conference', Antwerp, Belgium.

Hang, C., Kan, M.-Y. & Chua, T.-S. (2005), Generic soft pattern models for definitional question answering, *in* 'Proceedings of the 28th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR–2005)', Salvador, Brazil, pp. 384–391.

Harabagiu, S. (2004), Incremental topic signatures, *in* 'Proceedings of the 20th International on Computational Linguistics (COLING–2004)', Geneva, Switzerland, pp. 583–589.

Harabagiu, S. & Lăcătuşu, F. (2005), Topic themes for multi-document summarization, *in* 'Proceedings of the 28th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR-2005)', Salvador, Brazil, pp. 202–209.

Harabagiu, S. & Maiorano, S. (2002), Multi-document summarization with GISTexter, *in* 'Proceedings of the Third International Conference on Language Resources and Evaluation (LREC–2002)', Canary Islands, Spain.

Harabagiu, S., Moldovan, D., Clark, C., Bowden, M., Hickl, A. & Wang, P. (2005), Employing two question answering systems in TREC-2005, *in* 'Proceedings of the Fourteenth Text REtrieval Conference (TREC–2005)', Gaithersburg, Maryland, USA.

Harabagiu, S., Moldovan, D., Paşca, M., Mihalcea, R., Surdeanu, M., Bunescu, R., Gîrju, R., Rus, V. & Morărescu, P. (2001), The role of lexico-semantic feedback in open-domain textual question-answering, *in* 'Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics (ACL–2001)', Toulouse, France, pp. 282–289.

Harman, D. & Voorhees, E., eds (2001), *Proceedings of the First Document Understanding Conference (DUC–2001)*, NIST, New Orleans, Louisiana, USA.

Hasegawa, T., Sekine, S. & Grishman, R. (2004), Discovering relations among named entities from large corpora, *in* 'Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL–2004)', Barcelona, Spain, pp. 415–422.

Hatzivassiloglou, V., Klavans, J. L., Holcombe, M. L., Barzilay, R., Kan, M.-Y. & McKeown, K. R. (2001), SimFinder: A flexible clustering tool for summarization, *in* 'Proceedings of the NAACL–2001 Workshop on Automatic Summarization', Pittsburg, Pennsylvania, USA, pp. 41–49.

Hearst, M. (1992), Automatic acquisition of hyponyms from large text corpora, *in* 'Proceedings of the 14th conference on Computational linguistics (COLING–1992)', Nantes, France, pp. 539–545.

Hensler, C. & Stipak, B. (1979), 'Estimating interval scale values for survey item response categories', *American Journal of Political Science* **23**(3), 627–649.

Herman, D., Jahn, M. & Ryan, M.-L., eds (2005), *Routledge Encyclopedia of Narrative Theory*, Routledge, London, UK.

Hobbs, J. & Israel, D. (1994), Principles of template design, *in* 'Proceedings of Human Language Technology Workshop', Plainsboro, New Jersey, USA.

Hochbaum, D. S. (1997), Approximating covering and packing problems: Set cover, vertex cover, independent set, and related problems, *in* D. S. Hochbaum, ed., 'Approximation Algorithms for NP-hard Problems', PWS Publishing Company, Boston, Massachusetts, USA, pp. 94–143.

Hovy, E., Ide, N., Frederking, R., Mariani, J. & Zampolli, A. (1999), 'Multilingual information management: Current levels and future abilities'.

Hovy, E. & Lin, C.-Y. (1998), Automated text summarization in SUMMARIST, *in* M. Maybury & I. Mani, eds, 'Advances in Automatic Text Summarization', Cambridge: MIT Press.

Hovy, E., Marcus, M., Palmer, M., Pradhan, S., Ramshaw, L. & Weischedel, R. (2006), Ontonotes: The 90% solution, *in* 'Proceedings of the Joint Human Language Technology Conference and North American Chapter of the Association for Computational Linguistics Meeting (HLT/NAACL–2006)', New York, New York, USA.

Huttunen, S., Yangarber, R. & Grishman, R. (2002), Complexity of event structure in IE scenarios, *in* 'Proceedings of the International Conference on Computational Linguistics (COLING–2002)', Taipei, Taiwan, pp. 1–7.

Ipeirotis, P., Gravano, L. & Sahami, M. (2001), Probe, count, and classify: Categorizing hidden-web databases, *in* 'Proceedings of the ACM SIGMOD International Conference on Management of Data (SIGMOD–2001)', Santa Barbara, California, USA, pp. 67–78.

Jijkoun, V., Mur, J. & de Rijke, M. (2004), Information extraction for question answering: Improving recall through syntactic patterns, *in* 'Proceedings of the Conference on Computational LInguistics (COLING–2004)', Geneva, Switzerland, pp. 1284–1290.

Jones, K. S. (1993), Discourse modelling for automatic summarising, Technical report no. 290, University of Cambridge, Computer Laboratory.

Katz, B., Borchardt, G. & Felshin, S. (2005), Syntactic and semantic decomposition strategies for question answering from multiple resources, *in* 'Proceedings of the AAAI 2005 Workshop on Inference for Textual Question Answering', Pittsburgh, Pennsylvania, USA, pp. 35–41.

Katz, B., Marton, G., Borchardt, G., Brownell, A., Felshin, S., Loreto, D., Louis-Rosenberg, J., Lu, B., Mora, F., Stiller, S., Uzuner, O. & Wilcox, A. (2005), External knowledge sources for question answering, *in* 'Proceedings of the Fourteenth Text REtrieval Conference (TREC–2005)', Gaithersburg, Maryland, USA.

Kemeny, J. G. & Snell, J. L. (1960), *Finite Markov Chains*, Princeton, NJ: Van Nostrand.

Kim, J.-T. & Moldovan, D. (1995), 'Acquisition of linguistic patterns for knowledge-based information extraction', *IEEE Transactions on Knowledge and Data Engineering* **7**(5), 713–724.

King, G. & Lowe, W. (2003), 'An automated information extraction tool for international conflict data with performance as good as human coders: A rare events evaluation design', *International Organization* **57(3)**, 617–642.

Kingsbury, P. & Palmer, M. (2002), From treeBank to propBank, *in* 'Proceedings of the Third International Conference on Language Resources and Evaluation (LREC–2002)', Las Palmas, Canary Islands, Spain.

Klein, D. & Manning, C. (2002), Fast exact inference with a factored model for natural language parsing, *in* 'Proceedings of Advances in Neural Information Processing Systems 15 (NIPS–2002)', Vancouver, Canada.

Kleinberg, J. (1999), 'Authoritative sources in a hyperlinked environment', *Journal of the ACM* **46**(5), 604–632.

Koller, D. & Sahami, M. (1997), Hierarchically classifying documents using very few words, *in* 'Proceedings of the 14th International Conference on Machine Learning (ICML–1997)', Nashville, Tennessee, USA, pp. 170–178.

Kupiec, J. (1993), MURAX: A robust linguistic approach for question answering using an on-line encyclopedia, *in* 'Research and Development in Information Retrieval', pp. 181–190.

Kupiec, J., Pedersen, J. & Chen, F. (1995), A trainable document summarizer, *in* 'Proceedings of the 18th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR–1995)', Seattle, Washington, USA, pp. 68–73.

Kwok, K., Grunfeld, L., Dinstl, N. & Chan, M. (2000), TREC-9 cross language, web and question-answering track, *in* 'Proceedings of the Ninth Text REtrieval Conference (TREC–9)', pp. 26–35.

Lehnert, W. (1978), *The Process of Question Answering. A Computer Simulation of Cognition*, Lawrence Erlbaum Associates.

Lehnert, W., Cardie, C., Fisher, D., McCarthy, J., Riloff, E. & Soderlan, S. (1994), University of massachusetts: Description of the CIRCUS system as used for MUC-4, *in* 'Proceedings of the Fourth Message Understanding Conference (MUC–4)'.

Levin, B. (1993), *English Verb Classes and Alternations: A Preliminary Investigation*, The University of Chicago, Chicago, Illinois, USA.

Light, M., Mann, G., Riloff, E. & Breck, E. (2001), 'Analyses for Elucidating Current Question Answering Technology', *Journal for Natural Language Engineering* **7**(4), 325–342.

Lin, C.-Y. & Hovy, E. (1997), Identifying topic by position, *in* 'Proceedings of the 5th Conference on Applied Natural Language Processing (ANLP–1997)', Washington, DC, USA, pp. 283–290.

Lin, C.-Y. & Hovy, E. (2000), The automated acquisition of topic signatures for text summarization, *in* 'Proceedings of the 18th International Conference on Computational Linguistics (COLING–2000)', Saarbrücken, Germany, pp. 495–501.

Lin, C.-Y. & Hovy, E. (2003), Automatic evaluation of summaries using n-gram cooccurrence statistics, *in* 'Proceedings of the Joint Human Language Technology Conference and North American chapter of the Association for Computational Linguistics Annual Meeting (HLT/NAACL–2003)', Edmonton, Canada, pp. 71–78.

Lin, J. (2007), 'An exploration of the principles underlying redundancy-based factoid question answering', *ACM Transactions on Information Systems (TOIS)* **25**(2).

Lindsay, R. K. (1963), Inferential memory as the basis of machines which understand natural language, *in* E. A. Feigenbaum & J. Feldman, eds, 'Computers and Thought', McGraw-Hill, New York, pp. 217–236.

Lita, L. V., Hunt, W. & Nyberg, E. (2004), Resource analysis for question answering, *in* 'Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL–2004)', Barcelona, Spain, pp. 35–41.

Luhn, H. (1957), 'A statistical approach to mechanized encoding and searching of literary information', *IBM Journal of Research and Development* **1**, 309–317.

Magnini, B., Negri, M., Prevete, R. & Tanev, H. (2002), Is it the right answer? Exploiting web redundancy for answer validation, *in* 'Proceedings of the 40th Annual ACL Meeting', Philadelphia, Pennsylvania, USA, pp. 425–432.

Makkonen, J. (2003), Investigations on event evolution in tdt, *in* 'Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology (NAACL/HLT)', Edmonton, Canada, pp. 43–48.

Mann, G. S. (2002), Fine-grained proper noun ontologies for question answering, *in* 'Proceedings of the COLING–2002 Workshop on SEMANET: building and using semantic networks', Taipei, Taiwan, pp. 1–7.

Marcu, D. (1997), From discourse structures to text summaries, *in* 'Proceedings of the ACL/EACL–1997 Workshop on Intelligent Scalable Text Summarization', Madrid, Spain, pp. 82–88.

Marsh, E. & Perzanowski, D. (1997), MUC-7 evaluation of IE technology: Overview of results, *in* 'Proceedings of the Seventh Message Understanding Conference (MUC–7)'.

Maynard, D., Bontcheva, K. & Cunningham, H. (2003), Towards a semantic extraction of named entities, *in* 'Proceedings of the Conference on Recent Advances in Natural Language Processing (RANLP–2003)', Borovets, Bulgaria, pp. 255–261.

McCoy, K. F. & Strube, M. (1999), Taking time to structure discourse: Pronoun generation beyond accessibility, *in* 'Proceedings of the 21th Annual Conference of the Cognitive Science Society', Vancouver, Canada, pp. 378–383.

McDonald, R., Pereira, F., Kulick, S., Winters, S., Jin, Y. & White, P. (2005), Simple algorithms for complex relation extraction with applications to biomedical ie, *in* 'Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL–2005)', Ann Arbor, Michigan, USA, pp. 491–498.

McKeown, K. (1985), *Text Generation: Using Discourse Strategies and Focus Constraints to Generate Natural Language Text*, Cambridge University Press, Cambridge, England.

McKeown, K., Barzilay, R., Chen, J., Elson, D., Evans, D., Klavans, J., Nenkova, A., Schiffman, B. & Sigelman, S. (2003), Columbia's NewsBlaster: New features and future directions, *in* 'Proceedings of the Joint Human Language Technology Conference and North American Chapter of the Association for Computational Linguistics Meeting (HLT/NAACL–2003)', Edmonton, Canada, pp. 15–16.

McKeown, K. R. & Radev, D. R. (1995), Generating summaries of multiple news articles, *in* 'Proceedings of the 18th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR–1995)', Seattle, Washington, USA, pp. 74–82.

Meyers, A., Grishman, R., Kosaka, M. & Zhao, S. (2001), Covering treebanks with GLARF, *in* 'Proceedings of the ACL/EACL Workshop on Sharing Tools and Resources for Research and Education', Toulouse, France, pp. 51–58.

Miller, G. (1995), 'WordNet: a lexical database for English', *Communications of the ACM* **38**(11), 39–41.

Moldovan, D., Harabagiu, S., Gîrju, R., Morărescu, P., Lăcătuşu, F., Novischi, A., Badulescu, A. & Bolohan, O. (2002), LCC tools for question answering, *in* '11th Text Retrieval Conference (TREC–2002)', Gaithersburg, Maryland, USA.

Molina, P. (1995), 'Documentary abstracting: Towards a methodological model', *Journal of the American Society for Information Science* **46**(3), 225–234.

Mollá, D. & van Zaanen, M. (2005), Learning of graph rules for question answering, *in* 'Proceedings of the Australasian Language Technology Association (ALTW–2005)', Sydney, Australia.

Muslea, I. (1999), Extraction patterns for information extraction tasks: A survey, *in* 'Proceedings of AAAI Workshop on Machine Learning for Information Extraction', Orlando, Florida, USA.

Narayanan, S. & Harabagiu, S. (2004), Question answering based on semantic structures, *in* 'Proceedings of the 20th International on Computational Linguistics (COLING–2004)', Geneva, Switzerland, pp. 693–701.

Naughton, M., Kushmerick, N. & Carthy, J. (2006), Event extraction from heterogeneous news sources, *in* 'Proceedings of the AAAI 2006 Workshop on Event Extraction and Synthesis', Boston, Massachusetts, USA.

Nenkova, A. & McKeown, K. (2003), References to named entities: A corpus study, *in* 'Proceedings of the Joint Human Language Technology Conference and North

American chapter of the Association for Computational Linguistics Annual Meeting (HLT/NAACL–2003)', Edmonton, Canada.

Ono, K., Sumlta, K. & Miike, S. (1994), Abstract generation based on rhetorical structure extraction, *in* 'Proceedings of the International Conference on Computational Linguistics (COLING–1994)', Kyoto, Japan, pp. 344–348.

Onyshkevych, B. (1993), Template design for information extraction system, *in* 'Proceedings of the Fifth Message Understanding Conference (MUC–5)', Morgan Kaufmann.

Paşca, M., Lin, D., Bigham, J., Lifchits, A. & Jain, A. (2006), Names and similarities on the web: Fact extraction in the fast lane, *in* 'Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (COLING/ACL–2006)', Sydney, Australia, pp. 809–816.

Palmer, M., Gildea, D. & Kingsbury, P. (2005), 'The Proposition Bank: An annotated corpus of semantic roles', *Computational Linguistics* **31**(1), 71–106.

Pan, F., Mulkar, R. & Hobbs, J. (2006*a*), An annotated corpus of typical durations of events, *in* 'Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC–2006)', Genoa, Italy.

Pan, F., Mulkar, R. & Hobbs, J. (2006*b*), Learning event durations from event descriptions, *in* 'Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (COLING/ACL–2006)', Sydney, Australia, pp. 393–400.

Peng, F., Weischedel, R. M., Licuanan, A. & Xu, J. (2005), Combining deep linguistics analysis and surface pattern learning: A hybrid approach to chinese definitional question answering, *in* 'Proceedings of the Joint Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP–2005)', Vancouver, Canada, pp. 307–314.

Pradhan, S., Ward, W., Hacioglu, K., Martin, J. & Jurafsky, D. (2004), Shallow semantic parsing using support vector machines, *in* 'Proceedings of the Joint Human Language

Technology Conference and North American chapter of the Association for Computational Linguistics Annual Meeting (HLT/NAACL–2004)', Boston, Massachusetts, USA, pp. 233–240.

Prager, J., Chu-Carroll, J., Brown, E. & Czuba, K. (2006), Question Answering by Predictive Annotation, *in* T. Strzalkowski & S. Harabagiu, eds, 'Advances in open-domain Question-Answering (Text, Speech and Language Technology)', number 32 *in* 'Text, Speech and Language Technology', Springer.

Prager, J., Chu-Carroll, J. & Czuba, K. (2004), Question answering using constraint satisfaction: QA-by-Dossier-with-Constraints, *in* 'Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL–2004)', Barcelona, Spain, pp. 574–581.

Proceedings of the Fifth Message Understanding Conference (MUC–5) (1993), Morgan Kaufmann, Baltimore, Maryland, USA.

Proceedings of the Fourth Message Understanding Conference (MUC–4) (1992), Morgan Kaufmann.

Proceedings of the Sixth Message Understanding Conference (MUC–6) (1995), Morgan Kaufmann, Columbia, Maryland, USA.

Pustejovsky, J. (2000), Events and the semantics of opposition, *in* C. Tenny & J. Pustejovsky, eds, 'Events as Grammatical Objects', CSLI Publications, pp. 445–482.

Pustejovsky, J., Castaño, J., Ingria, R., Saurí, R., Gaizauskas, R., Setzer, A., Katz, G. & Radev, D. (2003), TimeML: A specification language for temporal and event expressions, *in* 'Proceedings of the International Workshop of Computational Semantics (IWCS–2003)', Tilburg, The Netherlands.

Pustejovsky, J., Hanks, P., Saurí, R., See, A., Gaizauskas, R., Setzer, A., Radev, D., Sundheim, B., Day, D., Ferro, L. & Lazo, M. (2003), The TimeBank corpus, *in* 'Proceedings of the Corpus Linguistics', Lancaster, UK.

Qiu, L., Kan, M.-Y. & Chua, T.-S. (2006), Paraphrase recognition via dissimilarity significance classification, *in* 'Proceedings of the Conference on Empirical Methods for Natural Language Processing (EMNLP–2006)', Sydney, Australia, pp. 18–26.

Radev, D. & McKeown, K. (1998), 'Generating natural language summaries from multiple on-line sources', *Computational Linguistics* **24**(3), 469–500.

Radev, D., Prager, J. & Samn, V. (2000), Ranking suspected answers to natural language questions using predictive annotation, *in* 'Proceedings of the 6th Applied Natural Language Processing Conference (ANLP–2000)', Seattle, Washington, USA, pp. 150–157.

Ravichandran, D. & Hovy, E. (2002), Learning surface text patterns for a question answering system, *in* 'Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL-2002)', Philadelphia, Pennsylvania, USA, pp. 41–47.

Riloff, E. (1993), Automatically constructing a dictionary for information extraction tasks, *in* 'Proceedings of the Eleventh National Conference on Artificial Intelligence (AAAI-1993)', Raleigh, North Carolina, USA, pp. 811–816.

Riloff, E. (1996), Automatically generating extraction patterns from untagged text, *in* 'Proceedings of the Thirteenth National Conference on Artificial Intelligence (AAAI-1996)', Cambridge, Massachusetts, USA, pp. 1044–1049.

Riloff, E. & Schmelzenbach, M. (1998), An empirical approach to conceptual case frame acquisition, *in* 'Proceedings of the 6th Workshop on Very Large Corpora', Montreal, Canada, pp. 49–56.

Riloff, E. & Shepherd, J. (1997), A corpus-based approach for building semantic lexicons, *in* 'Proceedings of the Second Conference on Empirical Methods in Natural Language Processing (EMNLP–1997)', Providence, Rhode Island, USA, pp. 117–124.

Rosario, B. & Hearst, M. (2004), Classifying semantic relations in bioscience text, *in* 'Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL–2004)', Barcelona, Spain, pp. 430–437.

Salton, G. (1971), *The SMART retrieval system*, Prentice-Hall, NJ.

Salton, G., Singhal, A., Mitra, M. & Buckley, C. (1997), 'Automatic text structuring and summarization', *Information Processing and Management* **33**(2), 193–207.

Saurí, R., Verhagen, M. & Pustejovsky, J. (2005), Evita: A robust event recognizer for QA systems, *in* 'Proceedings of the Joint Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP–2005)', Vancouver, Canada, pp. 700–707.

Schank, R. (1975), Conceptual dependency theory, *in* R. Schank, ed., 'Conceptual Information Processing', North-Holland and Elsevier, Amsterdam and New York, pp. 22–82.

Schank, R. (1982), *Dynamic Memory*, Cambridge Univ. Press.

Schank, R. & Abelson, R. (1977), *Scripts, Plans, Goals, and Understanding An Inquiry Into Human Knowledge Structures*, Lawrence Erlbaum.

Schiffman, B., Mani, I. & Concepcion, K. (2001), Producing biographical summaries: combining linguistic knowledge with corpus statistics, *in* 'Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics (ACL–2001)', Toulouse, France, pp. 458–465.

Schiffman, B., McKeown, K., Grishman, R. & Allan, J. (2007), Question answering using integrated information retrieval and information extraction, *in* 'Proceedings of the Joint Human Language Technology Conference and North American Chapter of the Association for Computational Linguistics Meeting (HLT/NAACL–2007)', Rochester, New York, USA, pp. 532–539.

Sekine, S. (2006), On-demand information extraction, *in* 'Proceedings of the Joint Conference of the International Committee on Computational Linguistics and the Association for Computational Linguistics (ACL/COLING-2006)', Sydney, Australia, pp. 731–738.

Shinyama, Y. & Sekine, S. (2006), Preemptive information extraction using unrestricted relation discovery, *in* 'Proceedings of the Joint Human Language Technology Conference and North American Chapter of the Association for Computational Linguistics Meeting (HLT/NAACL–2006)', New York, New York, USA, pp. 304–311.

Siegel, E. V. (1999), Corpus-based linguistic indicators for aspectual classification, *in* 'Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL–1999)', College Park, Maryland, USA, pp. 112–119.

Siegel, E. V. & McKeown, K. R. (2000), 'Learning methods to combine linguistic indicators: Improving aspectual classification and revealing linguistic insights', *Computational Linguistics* **26**(4), 595–628.

Simmons, R. (1965), 'Answering English questions by computer: A survey', *Communications of the ACM* **8**(1), 53–70.

Simmons, R., Klein, S. & McConlogue, K. (1964), 'Indexing and dependency logic for answering english questions', *Amer. Documentation* **15**(3), 196–204.

Soubbotin, M. & Soubbotin, S. (2001), Patterns and potential answer expressions as clues to the right answers, *in* 'Proceedings of the Tenth Text REtrieval Conference (TREC–10)', Gaithersburg, Maryland, USA, pp. 175–182.

Srihari, R. & Li, W. (2000), Question answering system supported by information extraction, *in* 'Proceedings of the 1st Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL–2000)', Seattle, Washington, pp. 166–172.

Stoyanov, V., Cardie, C. & Wiebe, J. (2005), Multi-perspective question answering using the OpQA corpus, *in* 'Proceedings of the Joint Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP–2005)', Vancouver, Canada, pp. 923–930.

Strohman, T., Metzler, D., Turtle, H. & Croft, W. B. (2005), Indri: A language-model based search engine for complex queries (extended version), Technical report ir-407, University of Massachusetts Amherst, Department of Computer Science, Center for Intelligent Information Retrieval.

Sudo, K., Sekine, S. & Grishman, R. (2003), An improved extraction pattern representation model for automatic ie pattern acquisition, *in* 'Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL–2003)', Sapporo, Japan, pp. 224–231.

Teufel, S. & Moens, M. (1997), Sentence extraction as a classification task, *in* 'Proceedings of the ACL/EACL–1997 Workshop on Intelligent Scalable Text Summarizaion', Madrid, Spain, pp. 58–65.

Thadani, K. (2007), Reducing document-wide redundancy through dependency tree alignment, Master's thesis, Columbia University, Department of Computer Science.

Thorne, J. (1962), *Thorne Automatic language analysis*, Arlington, Virginia, USA.

Turing, A. M. (1950), 'Computing machinery and intelligence', *Mind* **59**(236), 433–460.

Voorhees, E. M. (2001), Overview of the TREC-2001 Question Answering Track, *in* 'Proceedings of the Tenth Text REtrieval Conference (TREC–2001)', Gaithersburg, Maryland, USA.

Voorhees, E. M. (2002), Overview of the TREC-2002 Question Answering Track, *in* 'Proceedings of the Eleventh Text REtrieval Conference (TREC–2002)', Gaithersburg, Maryland, USA.

Voorhees, E. M. (2003*a*), Evaluating answers to definition questions, *in* 'Proceedings of the Joint Human Language Technology Conference and North American chapter of the Association for Computational Linguistics Annual Meeting (HLT/NAACL–2003)', Edmonton, Canada, pp. 181–188.

Voorhees, E. M. (2003*b*), Overview of the TREC-2003 Question Answering Track, *in* 'Proceedings of the Twelfth Text REtrieval Conference (TREC–2003)', Gaithersburg, Maryland, USA.

Voorhees, E. M. (2004), Overview of the TREC-2004 Question Answering Track, *in* 'Proceedings of the Thirteenth Text REtrieval Conference (TREC–2004)', Gaithersburg, Maryland, USA.

Voorhees, E. M. (2005), Overview of the TREC-2005 Question Answering Track, *in* 'Proceedings of the Fourteenth Text REtrieval Conference (TREC–2005)', Gaithersburg, Maryland, USA.

Voorhees, E. M. & Harman, D. (1999), Overview of the Eighth Text REtrieval Conference (TREC-8)., *in* 'Proceedings of the Eighth Text REtrieval Conference (TREC–8)', Gaithersburg, Maryland, USA.

Voorhees, E. M. & Harman, D. (2000), Overview of the Ninth Text REtrieval Conference (TREC-9)., *in* 'Proceedings of the Ninth Text REtrieval Conference (TREC–9)', Gaithersburg, Maryland, USA.

Weischedel, R., Xu, J. & Licuanan, A. (2004), Hybrid approach to answering biographical questions, *in* M. Maybury, ed., 'New Directions In Question Answering', AAAI Press, chapter 5, pp. 59–70.

White, M., Korelsky, T., Cardie, C., Ng, V., Pierce, D. & Wagstaff, K. (2001), Multi-document summarization via information extraction, *in* 'Proceedings of the First International Conference on Human Language Technology Research (HLT–2001)', San Diego, California, USA, pp. 1–7.

Woods, W. A. (1978), Semantics and quantification in natural language question answering, *in* M. Yovits, ed., 'Advances in Computers', Vol. 17, Academic Press, pp. 2–64.

Xu, J., Licuanan, A. & Weischedel, R. (2003), TREC-2003 QA at BBN: Answering definitional questions, *in* 'Proceedings of the Twelfth Text REtrieval Conference (TREC–2003)', Gaithersburg, Maryland, USA.

Yang, H., Chua, T.-S., Wang, S. & Koh, C.-K. (2003), Structured use of external knowledge for event-based open domain question answering, *in* 'Proceedings of the 26th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR–2003)', Toronto, Canada, pp. 33–40.

Yang, Y., Carbonell, J., Brown, R., Price, T., Archibald, B. & Liu, X. (1999), 'Learning approaches for detecting and tracking news events', *IEEE Intelligent Systems: Special Issue on Applications of Intelligent Information Retrieval* **14**(4), 32–43.

Yangarber, R. (2003), Counter-training in discovery of semantic patterns, *in* 'Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL–2003)', Sapporo, Japan, pp. 343–350.

Yu, H. & Hatzivassiloglou, V. (2003), Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences, *in* 'Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP–2003)', Sapporo, Japan, pp. 129–136.

Zaki, M. (2002), Efficiently mining frequent trees in a forest, *in* 'Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (SIGKDD–2002)', Edmonton, Canada, pp. 1021–1035.

Zhao, Y. & Karypis, G. (2001), Criterion functions for document clustering: Experiments and analysis, Technical Report 01-40, Department of Computer Science, University of Minnesota.

Zhou, L., Ticrea, M. & Hovy, E. (2004), Multi-document biography summarization, *in* 'Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP–2004)', Barcelona, Spain, pp. 434–441.

Appendix A

MUC-4. Template Example for "Terrorist Attack" Domain

A.1 General Template for Terrorist Attacks Domain

0. MESSAGE: ID
1. MESSAGE: TEMPLATE
2. INCIDENT: DATE
3. INCIDENT: LOCATION
4. INCIDENT: TYPE
5. INCIDENT: STAGE OF EXECUTION
6. INCIDENT: INSTRUMENT ID
7. INCIDENT: INSTRUMENT TYPE
8. PERP: INCIDENT CATEGORY
9. PERP: INDIVIDUAL ID
10. PERP: ORGANIZATION ID
11. PERP: ORGANIZATION CONFIDENCE
12. PHYS TGT: ID
13. PHYS TGT: TYPE
14. PHYS TGT: NUMBER
15. PHYS TGT: FOREIGN NATION
16. PHYS TGT: EFFECT OF INCIDENT
17. PHYS TGT: TOTAL NUMBER
18. HUM TGT: NAME
19. HUM TGT: DESCRIPTION
20. HUM TGT: TYPE
21. HUM TGT: NUMBER
22. HUM TGT: FOREIGN NATION
23. HUM TGT: EFFECT OF INCIDENT
24. HUM TGT: TOTAL NUMBER

Figure A.1: MUC–4 template for *Terrorist Attacks* domain

A.2 Document TST4-MUC4-0004

TST4-MUC4-0004

Santiago, 25 Aug 88 (Radio Chilena Network)

The military court upheld the indictment of Felipe Sandoval, youth sector president of the Christian Democratic Party, who was tried for offending Carabineros. The military court also decided to release Sandoval on bail. He was arrested on 17 August. The court decision was 3 to 1.

Also today, Military Prosecutor Fernando Torres Silva met with Supreme Court President Luis Maldonado for 30 minutes to discuss several issues related to cases under Torres' jurisdiction, such as the attack on General Pinochet and the kidnapping of Colonel Carreno.

It was learned that Torres officially submitted a request for the extradition of three Chileans arrested near Buenos Aires, who are reportedly linked to the kidnapping of Col Carreno. The Supreme Court must now process that request.

A.3 Key for TST4-MUC4-0004

0. MESSAGE: ID TST4-MUC4-0004
1. MESSAGE: TEMPLATE *
2. INCIDENT: DATE *
3. INCIDENT: LOCATION *
4. INCIDENT: TYPE *
5. INCIDENT: STAGE OF EXECUTION *
6. INCIDENT: INSTRUMENT ID *
7. INCIDENT: INSTRUMENT TYPE *
8. PERP: INCIDENT CATEGORY *
9. PERP: INDIVIDUAL ID *
10. PERP: ORGANIZATION ID *
11. PERP: ORGANIZATION CONFIDENCE *
12. PHYS TGT: ID *
13. PHYS TGT: TYPE *
14. PHYS TGT: NUMBER *
15. PHYS TGT: FOREIGN NATION *
16. PHYS TGT: EFFECT OF INCIDENT *
17. PHYS TGT: TOTAL NUMBER *
18. HUM TGT: NAME *
19. HUM TGT: DESCRIPTION *
20. HUM TGT: TYPE *
21. HUM TGT: NUMBER *
22. HUM TGT: FOREIGN NATION *
23. HUM TGT: EFFECT OF INCIDENT *
24. HUM TGT: TOTAL NUMBER *

Figure A.2: Filled-in tempalte for document TST4-MUC4-0004

A.4 Document TST4-MUC4-0006

TST4-MUC4-0006

San Salvador, 25 Aug 88 (Canal Doce Television)

A corporal and a guerrilla have been killed in the most recent clashes between the Army and the FMLN. According to military reports, Corporal Alexander Molina Granados was killed as he was trying to defuse a mine placed by the insurgents on La Noria Bridge, in San Marcos Lempa. According to military authorities, a lieutenant and five civilians who were passing through the area at the time were injured by the explosion.

Meanwhile, troops of the Civiplan Battalion killed an FMLN guerrilla in a skirmish on the slopes of La Cruz Hill in Chalatenango Department.

Residents of Quetzaltepeque reported that the rebels downed an electric tower last night on the outskirts of the town.

A.5 Key for TST4-MUC4-0006

0. MESSAGE: ID – TST4-MUC4-0006

1. MESSAGE: TEMPLATE – 1

2. INCIDENT: DATE – (25 AUG 88) / (24 AUG 88 - 25 AUG 88)

3. INCIDENT: LOCATION – EL SALVADOR: SAN MARCOS LEMPA (CITY)

4. INCIDENT: TYPE – BOMBING

5. INCIDENT: STAGE OF EXECUTION – ACCOMPLISHED

6. INCIDENT: INSTRUMENT ID – ”MINE”

7. INCIDENT: INSTRUMENT TYPE – MINE: ”MINE”

8. PERP: INCIDENT CATEGORY – TERRORIST ACT

9. PERP: INDIVIDUAL ID – ”INSURGENTS”

10. PERP: ORGANIZATION ID – ”FMLN”

11. PERP: ORGANIZATION CONFIDENCE – reported as fact: ”FMLN”

12. PHYS TGT: ID – ”LA NORIA BRIDGE”

13. PHYS TGT: TYPE – TRANSPORTATION ROUTE: ”LA NORIA BRIDGE”

14. PHYS TGT: NUMBER – 1: ”LA NORIA BRIDGE”

15. PHYS TGT: FOREIGN NATION – *

16. PHYS TGT: EFFECT OF INCIDENT – *

17. PHYS TGT: TOTAL NUMBER – -

18. HUM TGT: NAME – ”ALEXANDER MOLINA GRANADOS”

19. HUM TGT: DESCRIPTION – ”CORPORAL”: ”ALEXANDER MOLINA GRANADOS”; ”LIEUTENANT”; ”CIVILIANS”

20. HUM TGT: TYPE – ACTIVE MILITARY: ”ALEXANDER MOLINA GRANADOS”; ACTIVE MILITARY: ”LIEUTENANT”; CIVILIAN: ”CIVILIANS”

21. HUM TGT: NUMBER – 1: ”ALEXANDER MOLINA GRANADOS”; 1: ”LIEUTENANT”; 5: ”CIVILIANS”

22. HUM TGT: FOREIGN NATION – *

23. HUM TGT: EFFECT OF INCIDENT – DEATH: ”ALEXANDER MOLINA GRANADOS”; INJURY: ”LIEUTENANT”; INJURY: ”CIVILIANS”

24. HUM TGT: TOTAL NUMBER – *

Figure A.3: Filled-in tempalte for document TST4-MUC4-0006

Appendix B

Answers to GALE Year 1 Go/No-Go Evaluation Template 1 Questions

B.1 Answers to GALE Year 1 Go/No-Go Evaluation Template 1 questions, Version 1 (Year 1 Go/No-Go system)

GNG001 LIST FACTS ABOUT EVENT [Vice President Dick Cheney's shooting of Harry Whittington]
source date constraint: 11 February 2006 – 15 February 2006
source constraint: broadcast speech documents

1. NOT RELEVANT: Klobuchar and Kennedy now regularly attack one another on just those points, with Kennedy aided in his critique by Bell, who refuses to be counted out and who has staked out detailed policy positions generally more liberal than those of Klobuchar.

GNG004 LIST FACTS ABOUT EVENT [The Hamas victory in Palestinian parliamentary elections]
source date constraint: 01 January 2006 – 01 May 2006

1. RELEVANT: Hamas landslide victory in Palestinian parliamentary elections last month January, 2006 prompted threats from the United States and European Union to cut off massive

aid to the Palestinians unless the group recognizes Israel and renounces violence.

2. RELEVANT: Russian officials promised they would demand that Hamas recognize the state of Israel and abandon the use of violence.

GNG005 LIST FACTS ABOUT EVENT [Appearance of human cases of bird flu in China]
source date constraint: 01 October 2005 – 01 May 2006

1. RELEVANT: Bird flu has ravaged Asian poultry stocks since late 2003, killing or forcing the slaughter of millions of birds.

2. NOT RELEVANT: The disease has apparently spread among local birds, not migratory birds, said Mohammed.

3. NOT RELEVANT: If those tests come back positive, Indonesias official human death toll from the virus would climb to 18.

4. RELEVANT: H5N1 can be deadly to humans, but no human deaths have been reported in either country.

5. NOT RELEVANT: Figures for quantities of imports were not immediately available, but EU is understood not to be a large amount.

6. NOT RELEVANT: A local zoological institute in the Foggia area of southern Italy, where the ducks were discovered dead, has sent the samples to national laboratories in the northern city of Padua to determine if the virus is the highly pathogenic form of H5N1, the ANSA news agency reported on Thursday 2/16/2006.

7. NOT RELEVANT: Families will drop off their child in the ICU and the next time they see them they will be dying.

8. NOT RELEVANT: It is likely that it will take a long time for the virus to be eradicated from the region, notwithstanding efforts by countries to cull and vaccinate their poultry populations, Satkunanatham said.

9. NOT RELEVANT: The EU commission spokesman Philip Tod, spokesman for EU health commissioner Markos Kyprianou. added that was assumed that the swans were following a migratory path further south than usual because of the particularly cold winter in Europe this year.

10. NOT RELEVANT: Dr. Ibtisam Aziz Ali, spokeswoman of a government committee dealing with the bird flu crisis, said Health Minister Abdel Mutalib Mohammed declared the alert after birds suspected of having avian influenza were discovered in at least five of parts of Maysan province, which borders Iran and lies on a major trade route between Basra and Baghdad.

11. RELEVANT: However Mao reassured there was no evidence of the virus mutating into a strain that could be transmitted easily among humans, circumstances that health experts fear would cause a global pandemic that would kill millions of people.

12. RELEVANT: Meanwhile, in Hong Kong on Wednesday 2/8/2006, government health officials said a dead egret found in a suburb of Hong Kong has tested positive for H5N1 strain.

13. RELEVANT: Earlier Wednesday 2/8/2006, China reported China 29th reported outbreak of the virus in birds since Oct. 19 10/19/2005.

14. NOT RELEVANT: The researchers did not use Japanese because it was the language the birds normally listened to, the newspaper said.

15. RELEVANT: Roche has signed sub-licensing agreements with two companies in China and India to produce Tamiflu, and is in talks with 12 other companies to boost production.

16. NOT RELEVANT: Specifically the European Union executive proposed the automatic agreement of protection zones as soon as a case of the H5 subtype of the virus is found, bolstering the process which until now has involved case-by-case decisions.

17. NOT RELEVANT: Azar, speaking to reporters, said the alliance would help prepare for any possible human flu pandemic and respond to those cases where bird flu has infected poultry and people.

18. NOT RELEVANT: Meanwhile, Christina Carlson, a U.N. specialist working in the former Soviet Union, said that experts had studied the capacities of hospitals in Crimea to deal with cases of human infection and found they have sufficient capacity to handle such cases if they will come up.

19. RELEVANT: Hong Kong authorities on Thursday 2/9/2006 investigated two new suspected cases of bird flu, a dead chicken and a wild bird found near a local school, and the health secretary said he wont be surprised if more birds fall ill.

20. RELEVANT: Up to 15,000 fowl in Yijing, a town in Chinas northern Shanxi province, were found dead late last week and tested positive for the H5N1 strain, Xinhua said.

21. NOT RELEVANT: The announcement came after the potentially deadly H5N1 strain of the disease was confirmed in swans in EU states Greece and Italy as well as EU hopeful Bulgaria, while the H5 strain was found in EU newcomer state Slovenia.

22. NOT RELEVANT: The virus has so far only been detected in birds passing through Hungary, and the aim is to prevent it from crossing over to local birds, said Gyurcsany, adding that should birds bred in Hungary be infected, vaccine would be available to protect people in close contact with them.

23. NOT RELEVANT: According to the European Union EU head office, the Slovenian authorities immediately applied the same precautionary measures as those set out in the Commission Decision adopted for Greece on Friday 2/10/2006 and applied in Italy during the weekend.

24. NOT RELEVANT: The alert is the latest measure taken by Iraqi health authorities to combat avian influenza H5N1 following last month January, 2006s discovery of the country Iraqis first and only confirmed bird flu case in a human.

25. RELEVANT: He said some patients could be living in environments that became contaminated with the virus through unknown channels, with the virus in those cases not necessarily causing large-scale deaths among animals.

26. NOT RELEVANT: Watanabe said paddy birds like the Java Sparrow and parakeets, which are skilled vocally, learn sounds unique to their species after becoming adults, suggesting that they have a high ability to distinguish between sounds.

GNG006 LIST FACTS ABOUT EVENT [Plots or attacks against US soldiers in Kuwait]
source date constraint: 01 January 2002 – 30 January 2006

1. NOT RELEVANT: A US truckdriver pleaded guilty to planning attacks against the United States with the al-Qaeda network as part of a deal with US authorities, Attorney General John Ashcroft announced Thursday 6/19/2003.

2. NOT RELEVANT: Hakim, whose Supreme Assembly of the Islamic Revolution in Iraq SAIRI was the main Shiite group SAIRI opposing Saddam, called for dialogue with the United States, moulding Iraqi public opinion to apply pressure and the setting up of an Iraqi administration to fill the political vacuum left by Saddams ouster.

3. NOT RELEVANT: Oteibi was arrested on the spot while militants in the car drove away and were being hunted by police the interior ministry, the ministry said.

4. NOT RELEVANT: On Thursday 5/29/2003, a U.S. soldier was killed when his convoy came under fire from a rocket-propelled grenade on a supply route through Iraq, bringing to nine the number of American soldiers who have died around the country the United States this week.

5. NOT RELEVANT: Al-Mutairi, who worked as a psychology researcher at the Ministry of Labor and Social Affairs, fled to Saudi Arabia after the killing. He was arrested there and extradited to Kuwait.

6. NOT RELEVANT: The Kuwait mission fell under Operation Enduring Freedom, launched against the al-Qaida terror network following the Sept. 11, 2001 attacks in the United States. But Germany said the unit would also have been ready to help if Iraq attacked Kuwait or U.S. forces there with weapons of mass destruction.

7. NOT RELEVANT: The militant, named as Fawaz Talaiq al-Oteibi, died in hospital after the shootout which erupted as security men came to arrest Oteibi in a Kuwait City suburb, said an interior ministry statement.

8. NOT RELEVANT: Kuwait has been a major Washington ally Kuwaiti since the 1991 U.S.-led Gulf War that liberated it from a seven-month Iraqi occupation.

9. NOT RELEVANT: A security source announced that two security men were killed and two others wounded in a shootout with gunmen in Hawally, state-run TV announced, without providing further details.

10. NOT RELEVANT: PARIS: The Paris prosecutors office said no expulsion orders would be issued against 17 detained Iranian opposition members, including figurehead leader Maryam Rajavi, as tension rose pending their appearance before anti-terrorist judges.

11. NOT RELEVANT: Plainclothes police tried to arrest Oteibi as Oteibi went to return a rented car in Hawalli, some 10 kilometers six miles south of Kuwait City.

12. NOT RELEVANT: And one American soldier was killed and another injured in an attack on a military convoy near the city of modern at a distance of 190 kilometers southeast of Baghdad.

13. NOT RELEVANT: One of the men in the car shot at the police killing two the interior ministry and wounding two others, the interior ministry said.

14. NOT RELEVANT: Rumsfeld said that for the first time the attacks against American troops were being coordinated regionally and possibly nationally by remnants of ousted Iraqi leader Saddam Husseins security forces, guerrilla fighters and Iraqi prisoners released before the war.

15. RELEVANT: The soldiers, some of them high-ranking officers, were detained and questioned a week ago over the plot, a Kuwaiti security source was quoted.

16. NOT RELEVANT: And the American helicopters over the area of the explosion the west bank of the Tigris River.

17. NOT RELEVANT: In response to the recent violence, McKiernan said McKiernan may soon send more troops into combat operations.

18. NOT RELEVANT: Some 12,000 American civilians live in Kuwait alongside around 9,000 Europeans and some 1,000 Australians.

19. RELEVANT: Kuwaiti security forces have detained up to eight soldiers suspected of planning to attack US forces in the emirate, the Arab Times reported Tuesday 1/4/2005.

20. NOT RELEVANT: The unit remained in place during the war in neighboring Iraq despite the German government Iraqs staunch opposition to U.S. military action and its refusal to contribute any troops.

21. NOT RELEVANT: The correspondent, Adel Eidan, Adel Eidan a Kuwaiti, reported Wednesday 1/5/2005 that two gunmen were arrested by Kuwaiti security forces after an exchange of fire with policemen in a suburb south of the capital.

22. RELEVANT: A Kuwaiti army spokesman said Thursday 1/6/2005 that two Kuwaiti soldiers were to stand trial on suspicion of plotting attacks on US and other foreign forces in the country, which served as the main launchpad for the war.

23. NOT RELEVANT: Earlier on Monday 1/10/2005, Kuwait began to ease tight security measures introduced almost two weeks ago when the emirate raised its state of alert almost to the maximum in the biggest show of force since the March 2003 March, 2003 launch of the US-led war in Iraq.

24. NOT RELEVANT: US citizens are urged to consider their safety and security before traveling to the Kyrgyz Republic, Kyrgyz Republic said.

25. NOT RELEVANT: Two other policemen were wounded in the firefight which came as the US embassy warned its nationals that militants at large in a car could randomly attack Westerners in the emirate.

26. NOT RELEVANT: KUWAIT CITY, Jan 10 1/10/2005 AFP - Two Kuwaiti security men were killed and two others wounded in a shootout with gunmen in a suburb south of the capital KUWAIT CITY on Monday 1/10/2005, Kuwait television reported.

27. NOT RELEVANT: The commander of coalition ground forces in Iraq said Thursday 5/29/2003 that recent attacks on U.S. forces were orchestrated by Baath Party groups loyal to ousted dictator Saddam Hussein.

28. NOT RELEVANT: Palacio, whose country currently holds the presidency of the UN Security Council, described the death of the expert David Kelly as an enormous tragedy which has highlighted the complexity of Iraqi weapons issue.

29. NOT RELEVANT: Kuwait in December December, 2004 raised its state of alert almost to the maximum, two weeks after the US embassy there warned US had credible information that terrorist groups were preparing to carry out attacks in the emirate in the near future.

30. NOT RELEVANT: Defense Minister Peter Struck did not specify when it would pull out.

31. NOT RELEVANT: It reminded American citizens of the potential for further terrorist actions against US citizens abroad, including in the Gulf region Gulf.

32. NOT RELEVANT: United States Defense Secretary Donald Rumsfeld said America might need to send additional forces to Iraq where more well-organized attacks against US soldiers are emerging, the New York Times reported Monday 7/14/2003.

33. NOT RELEVANT: The US embassy in Kuwait earlier Monday 1/10/2005 warned Kuwait citizens that assailants in a black car driving around Kuwait planned to randomly attack Westerners, in a message posted on the embassy website.

34. NOT RELEVANT: Kuwaiti Defense Minister Sheik Jaber Mubarak Al Hamad Al Sabah said Tuesday 1/4/2005 authorities had detained no more than three Kuwaiti armed forces personnel and added that we cant say if they are religious extremists.

35. NOT RELEVANT: I agreed with our American friends that we will bring our people home too.

36. NOT RELEVANT: Americans who encounter suspicious vehicles matching this description should quickly but safely move away and contact the Kuwaiti police emergency number at 777, it added.

37. RELEVANT: Around 25,000 US soldiers are stationed in Kuwait, which is also the main transit point for other coalition forces traveling to Iraq.

38. NOT RELEVANT: The explosion took place near the scene of on the west bank of the Tigris River, causing a crater in the street with a glass shattered the without having apparently injuring one.

39. RELEVANT: Most attacks were carried out by fundamentalist Muslims who do not approve of the U.S. military presence in Kuwait.

40. NOT RELEVANT: The arrests follow Kuwaits stepping up Kuwait internal security in recent days. Kuwait has stationed armed vehicles at street junctions, hotels, and embassies.

41. NOT RELEVANT: Two U.S. soldiers were killed and nine wounded there Sunday 5/25/2003 night during a firefight at a U.S. checkpoint in the town of 200,000 people Fallujah, known for supporting Saddam Hussein and Saddam Baath Party.

42. NOT RELEVANT: But Germany said the unit would also have been ready to help if Iraq attacked Kuwait or U.S. forces there with weapons of mass destruction.

43. NOT RELEVANT: The Kuwait army said Monday 1/3/2005 it has arrested a number of soldiers suspected of planning to attack US forces in the emirate, an army spokesman said.

44. NOT RELEVANT: KUWAIT CITY, Jan 10 1/10/2005 AFP - Kuwaiti security forces fatally wounded a wanted militant Monday 1/10/2005 in a deadly shootout in which two policemen were killed amid a US security alert, the interior ministry said.

45. NOT RELEVANT: The killing of two other Iraqis in falujah in an attack on the forces of the American Wednesday 5/21/2003 night.

46. NOT RELEVANT: Now that the Americans and Czechs have left, there is no longer a combined joint task force, Americans said.

47. NOT RELEVANT: Now that the 1st Armored Division has assumed the responsibility for the Baghdad area, Im working with the V Corps commander on different options, McKiernan said.

48. NOT RELEVANT: The falujah scene of clashes bloody end when 16 Iraqis by American soldiers during the demonstrations.

49. NOT RELEVANT: The statement added that the attackers who did not know their number, attacked the forces of the American missiles and with light weapons.

50. NOT RELEVANT: McKiernan said the attacks are being perpetrated by enemies whose future is gone... the rest of the population knows that they were thugs under McKiernan regime and they know, and the Iraqi population knows that they have no future in this country.

51. NOT RELEVANT: Staunch US ally Kuwait last week raised its state of alert almost to the maximum, boosting security around the country in the biggest show of force since the March 2003 March, 2003 launch of the US-led war in Iraq.

52. NOT RELEVANT: Kuwait is holding 32 suspects, including two women, in connection with a series of deadly gunfights in January January, 2005 between Islamist militants and security forces, the minister of justice said in remarks published Thursday 3/10/2005.

53. NOT RELEVANT: The statement went on to say that the kind of Midfak affiliated to the American forces suffered damage during the battle when an offer by Bradley had been to the ceasefire.

54. NOT RELEVANT: Police found a number of hand grenades, arms and ammunition in the gunmens car which had suspiciously circled state security headquarters in the area before the shooting erupted, the correspondent Adel Eidan said.

55. NOT RELEVANT: The foreign minister of Spain, one of the key US allies in the war in Iraq, urged patience Sunday 7/20/2003 over the search for Saddam Husseins alleged arsenal of banned arms, amid mounting doubts that such weapons will ever be found.

56. NOT RELEVANT: U.S. authorities will not disclose the exact figure of American forces in Kuwait, which was used by coalition forces to invade neighboring Iraq in March 2003 March, 2003.

57. NOT RELEVANT: The V Corps is an umbrella operation that coordinates American forces in Iraq.

58. NOT RELEVANT: The war has not ended, Lt. Gen. David McKiernan told reporters at a news conference. Decisive combat operations against military formations has ended, but these

contacts were having right now are in a combat zone, and it is war, and they are members of Saddams regime.

59. NOT RELEVANT: If we need to apply some of the combat power of the 3rd Infantry Division elsewhere in Iraq, we will certainly not hesitate to do that, McKiernan said.

60. NOT RELEVANT: With recent attacks against U.S. soldiers, McKiernan said there were no immediate plans to return the unit to Army headquarters at Fort Stewart, Georgia.

61. NOT RELEVANT: The embassy is issuing this urgent message because it has received credible information that an individual or individuals moving about Kuwait in a black coloured small sedan intend to randomly attack Westerners, the message said.

62. NOT RELEVANT: The embassy called on all US citizens to exercise caution, maintain a low profile, and avoid areas where Westerners are known to congregate.

63. NOT RELEVANT: McKiernan said the Armys 3rd Infantry Division, which had been planning to return to the United States in June June, 2003, was going to remain in Iraq until commanders decided they were no longer needed.

64. NOT RELEVANT: The Pentagon said the soldier was part of a convoy that was attacked north of Baghdad.

GNG025 LIST FACTS ABOUT EVENT [The shut down of the Cernavoda nuclear power plant]
location constraint: Romania
source date constraint: 24 August – 01 December 2003

1. NOT RELEVANT: The 152-page manual deals with Holocaust denial, figures for the number of Jews killed, details about concentration camps, death chambers, and the persecution of Gypsies, homosexuals and Jehovah Witnesses.

2. NOT RELEVANT: The book, called Teaching the Holocaust in the 21st century, was written by a French author and translated into Romanian.

3. NOT RELEVANT: Today 9/17/2003, about 6,000 Jews live in Romania.

4. NOT RELEVANT: Romania participated in the Holocaust and we have to face history, the minister, Razvan Teodorescu, said at a ceremony marking the inauguration of a teaching manual on the Holocaust.

5. NOT RELEVANT: The manual, which will be given to schools around Romania, was issued three months after the Romanian government Romanian said in a statement that there had been no Holocaust inside Romanias borders.

6. NOT RELEVANT: The barbarism of the Holocaust was unique in history and should not be repeated, Teodorescu said at the ceremony, which was attended by U.S. and Israeli diplomats as well as several representatives from the Jewish community.

7. NOT RELEVANT: Israel protested and Iliescu pledged to support Holocaust education in Romania and set up a memorial day to commemorate Holocaust victims.

8. NOT RELEVANT: Romania was a German ally Romanian during most of World War II and tens of thousands of Jews and Gypsies died in concentration camps.

9. NOT RELEVANT: Romania introduced a new program Wednesday 9/17/2003 to teach high school students about the Holocaust, marking the country Romanians first effort to educate Romania youth about Romania role in the deaths of thousands of Jews during World War II.

10. RELEVANT: A nuclear power plant was shut down Sunday 8/24/2003 because a record drought left insufficient water to cool down the reactor. The plant supplies more than 10 percent of Romania's electricity and closure prompted fears of a price hike.

11. NOT RELEVANT: During communist times, schoolbooks taught that Germans were the sole perpetrators of the Holocaust, ignoring the involvement of Romanias wartime leaders.

12. RELEVANT: Its the first time the plant in Cernavoda, some 200 kilometers 125 miles east of Bucharest, has encountered water level problems since Bucharest opened seven years ago.

13. NOT RELEVANT: Romania is undergoing economic transition in recent years, said the president Visiting Romanian President Ion Iliescu, adding Romania has taken measures to stimulate foreign investment, and welcomes Hong Kong businessmen to directly invest in the country Romanian.

14. NOT RELEVANT: Romanias traditionally good ties to Israel were strained in June June, 2003 after the government claimed there was no Holocaust inside Romanias borders. Romania was a German ally during most of World War II.

15. NOT RELEVANT: Since the fall of communism in 1989, schools have only slowly begun to address the issue of Romanian involvement in the atrocities.

16. NOT RELEVANT: Romanias culture minister urged his country Tuesday 9/16/2003 to fully acknowledge its role in killing thousands of Jews during the Holocaust, an apparent attempt to heal relations with Israel.

17. NOT RELEVANT: Visiting Danish Prime Minister Anders Fogh Rasmussen said Thursday 9/4/2003 that his country would continue to support Romania in Romania bid to join the European Union EU.

18. NOT RELEVANT: Romania and Hungary signed Tuesday 9/23/2003 an agreement acknowledging the validity of a Hungarian law expanding benefits to ethnic Hungarians in neighboring countries, including Romania.

19. NOT RELEVANT: But tensions resurfaced again in July July, 2003 when Romanian President Ion Iliescu was quoted by an Israeli paper as saying, the Holocaust was not unique to the Jewish population in Europe. Many others, including Poles.

20. NOT RELEVANT: Israel and Romanias Jewish community protested, and the government eventually acknowledged that Romanias wartime leaders deported and killed Jews.

GNG041 LIST FACTS ABOUT EVENT [a ferry crash] (location constraint: Canada)
location constraint: Canada
source date constraint: 01 February 2005 – 30 June 2005
source constraint: broadcast speech documents

1. NOT RELEVANT: November 29: Some 60 people drown after a passenger ferry goes down in the southern Mehendiganj coastal subdistrict.

2. NOT RELEVANT: The Afghan government Afghan has said the cause of the crash remains unknown and have called in U.S. experts to help investigate.

3. NOT RELEVANT: A catamaran ferry carrying about 156 passengers collided with a Chinese cargo vessel off Hong Kongs Tsing Yi Island, leaving four people seriously hurt, a government spokeswoman said.

4. NOT RELEVANT: May 3: Up to 200 people die when a double-decker ferry sinks near the southeastern Chandpur river port after being hit by a storm during the night.

5. NOT RELEVANT: November 29: Some 60 people drown after a passenger ferry goes down in the southern Mehendiganj coastal subdistrict. October 19 10/19/2004: Seven die after a

wooden ferry sinks in a waterway in the northeastern Sunamganj district. June 1 6/1/2005: Twenty-six people are killed when an overloaded ferry sinks off the southeastern coast near the port city of Chittagong.

6. NOT RELEVANT: April 23 4/23/2005: A passenger ferry sinks in the Buriganga river near Dhaka, leaving leaving at least 129 people dead. On the same day a second ferry carrying a bridal party sinks in northern Kishoreganj district killing 52. April 7 4/7/2005: At least 72 people die in a ferry tragedy in northeastern Sylhet district. April 1 4/1/2005: Thirteen people are killed when a river ferry sinks in western Bangladesh.

7. NOT RELEVANT: On the same day a second ferry carrying a bridal party sinks in northern Kishoreganj district killing 52.

8. NOT RELEVANT: Maj. Gen. Mohammed Moeen Faqir, an Afghan army commander, said the teams had not yet been able to recover any of the bodies. However, Defense Ministry spokesman Gen. Mohammed Zaher Azimi said the flight recorder had been found on Sunday 2/13/2005.

9. NOT RELEVANT: NATO and Afghan troops have retrieved the flight recorder from a crashed Afghan airliner, an Afghan official said, 10 days after the plane smashed into a mountain in a snowstorm, killing all 104 people on board. The first clear weather in nearly a week allowed helicopters to ferry troops and investigators to the crash site, 3,000 meters 10,000 feet up a snow-covered peak about 30 kilometers 20 miles east of the capital, officials said. Maj. Gen. Mohammed Moeen Faqir, an Afghan army commander, said the teams had not yet been able to recover any of the bodies. However, Defense Ministry spokesman Gen. Mohammed Zaher Azimi said the flight recorder had been found on Sunday 2/13/2005. It is in the hands of the investigating commission Defense Ministry, Azimi said. Azimi.

10. NOT RELEVANT: im told we have an eyewitness that will bring it to you thats that uh b. c. ferry crash that happened in west vancouver.

11. NOT RELEVANT: The Boeing 737 crashed into the mountaintop east of the capital, Kabul, on Feb. 3 2/3/2005 after approaching from the western city of Herat. Authorities have declared all 96 passengers and eight crew dead, including more than 20 foreigners, in the country Afghan.

12. NOT RELEVANT: A Canadian national was among the 104 people aboard an Afghan jet which crashed in the mountains east of Kabul earlier this week, the Canadian foreign ministry said

Saturday 2/5/2005.

13. NOT RELEVANT: 1987 April 14: Some 90 passengers die as a ferry sinks in the Meghna.

14. NOT RELEVANT: 1999 December 11: Fourty-six people are killed when a ferry sinks in the Meghna river near Chandpur. May 8 5/8/2005: A passenger ferry goes down in the Meghna near the coastal district of Bhola killing 72 coastal Pirojpur district.

15. NOT RELEVANT: Bucharest Television said bad weather was responsible for the crash. Since January 1994 January, 1994, eight Romanian helicopters have crashed, the Romanian News Agency Rompres.

16. NOT RELEVANT: July 9: At least 156 die when an overloaded ferry capsizes at the confluence of three rivers including the Meghna River near southeastern Chandpur.

17. NOT RELEVANT: Two passenger ferries collided with cargo boats in heavy fog in Hong Kong Thursday 2/17/2005, injuring more than 90 people, the government Hong Kong said.

18. RELEVANT: uh mark uh landry whos on that ferry that crashed into the marina at uh horseshoe bay in west vancouver.

19. NOT RELEVANT: April 7 4/7/2005: At least 72 people die in a ferry tragedy in northeastern Sylhet district.

20. NOT RELEVANT: They were all taken to hospital for treatment, she said.

21. NOT RELEVANT: June 1 6/1/2005: Twenty-six people are killed when an overloaded ferry sinks off the southeastern coast near the port city of Chittagong.

22. NOT RELEVANT: i dont know you know i actually did do a very good job what you keep in touch with uss because you could be a very good on the scene reporter mark im glad youre ok take care mark you are a that was mark landry alive from the ferry at ferry that crashed there the b. c. ferry uh.

23. NOT RELEVANT: Feb 4 2/4/2005: Twelve people die in a collision between two ferries in foggy weather in the southern Barisal district.

24. NOT RELEVANT: 1989 May 18: Double-decker ferry sinks in Meghna river killing 95 passengers.

25. RELEVANT: i would i think around and we travel the ferry probably ten or fifteen times a year so this is not a common occurrence to crash into the marina.

26. NOT RELEVANT: Bad weather had previously allowed only a brief inspection of the crash site, which is covered in deep snow, but NATO officials said a team of de-miners was able to spend four hours on Sunday 2/13/2005 making sure that a makeshift landing pad near an old military lookout on the summit was safe. Afghan officials said their troops planned to erect a tent to hold collected remains before they can be flown out by helicopter, though they have warned that the recovery operation could take several weeks. The Afghan government Afghan has said the cause of the crash remains unknown and have called in U.S. experts to help investigate. The Afghan transport minister has said the plane disappeared from radar screens shortly after it was cleared to land in Kabul, though the private airline, Kam Air, Kam Air says the pilot had turned away from the capital to seek an easier landing in Pakistan.

27. NOT RELEVANT: The capsizing of a ferry at the weekend, with 116 people confirmed dead by Monday 2/21/2005 and scores believed missing, is the latest in a string of ferry disasters that have killed thousands of people in Bangladesh.

28. NOT RELEVANT: A helicopter ferrying food to wild boars in the Carpathians went down in western Romania Friday 2/18/2005, with two of the four people aboard killed and the other two seriously injured.

29. NOT RELEVANT: May 2: At least 127 people die when two small ferries sink in storms in the eastern district of Brahmanbaria. March 14 3/14/2005: Twelve people die in a ferry crash in Bishkhali river in coastal Pirojpur district.

30. NOT RELEVANT: The capsizing of a ferry at the weekend, with 116 people confirmed dead by Monday 2/21/2005 and scores believed missing, is the latest in a string of ferry disasters that have killed thousands of people in Bangladesh. The country is criss-crossed by a network of 230 rivers and ferry travel is a way of life, with some 3,000 ferries providing transport for more than 100,000 people each day. Since 1977, more than 260 ferry accidents have claimed the lives of at least 3,000 people. The following are some of the other major ferry accidents.

31. NOT RELEVANT: A helicopter ferrying food to wild boars in the Carpathians crashed in bad weather in central Romania Friday 2/18/2005, killing two of the four people aboard and injuring the other two, regional governor Cristian Vladu said.

32. NOT RELEVANT: May 4: 170 people lose their lives in a major ferry crash in the Kalabadar river in coastal Patuakhali district.

33. NOT RELEVANT: May 23: Eighty-one people die when two separate boats capsize during a pre-monsoon storm on a stretch of river in southeastern Chandpur.

34. NOT RELEVANT: May 2: At least 127 people die when two small ferries sink in storms in the eastern district of Brahmanbaria.

35. NOT RELEVANT: January 22: Eighteen people die and 25 others go missing in a ferry capsize in Kacha river of Patuakhali. 1994 August 1: Thirty passengers die in a ferry crash in Paira river in Patuakhali.

GNG044 LIST FACTS ABOUT EVENT [hunger strikes by Palestinians in Israeli jails]
source date constraint: 01 August 2004 – 31 August 2004

1. RELEVANT: For its part, the Al-Aqsa Martyrs Brigade of the Fatah movement of the Palestinian Authority to provide free education to all prisoners in support of the cause of detainees in their hunger strike.

2. NOT RELEVANT: She added that a gun near the bodies of the dead, who was wearing a bullet-proof vest.

3. RELEVANT: Systems thousands of Palestinians massive demonstrations today 8/21/2004 in the cities of the West Bank and Gaza Strip, to express their support for the detainees Palestinians on hunger strike in Israeli jails.

4. RELEVANT: International Committee is noteworthy that three to four thousand Palestinian out of a total of 7500 in the prisons of the Israeli The fifteenth of August August, 2004 on hunger strike demanding an improvement in their conditions of detention.

5. NOT RELEVANT: The international organization the United Nations in a joint statement issued by more than ten subsidiaries operating in the territories of the occupied Palestinian Israel to Israel obligations in accordance with the fourth Geneva Convention on the protection of civilians in time of war and other human rights charters and which provided for the protection of detainees and prisoners.

6. RELEVANT: Today 8/24/2004, Tuesday 8/24/2004 club of Palestinian captive of Pope John Paul the second to intervene on the side of the detainees Palestinians on hunger strike in Israel.

7. NOT RELEVANT: It said that it is the third anniversary of the death of its Secretary General.

8. NOT RELEVANT: He added that it had been formed medical teams of security measures and avoid any chaos.

9. NOT RELEVANT: Command declared Palestinian national and Islamic to strike in the prisons of the Israeli Saturday 8/28/2004 afternoon commenting on what was published yesterday 8/27/2004 by the news about the suspension of hunger strike in prison that what happened in the not to strike but commenting.

10. RELEVANT: Officials said the Palestinians today 8/28/2004, Saturday 8/28/2004, that the resolution of the Palestinian prisoners in prison the Israeli suspension of their hunger strike until the day after tomorrow 8/29/2004, Monday 8/23/2004 came after the approval of the Israeli Prisons Department meet some of Prisons Department demands and the approval of the examination of the other demands.

11. NOT RELEVANT: We wish you after prayers and pray to protect our prisoners of oppression and terrorism the Israeli succeed offices in deterring kill human being pursued by the Government of Israel.

12. RELEVANT: Erakat said in an interview with the Voice of Palestine radio today 8/31/2004, I am ready to hold a meeting with Israeli officials to discuss the demands on the detainees.

13. NOT RELEVANT: And the demands of the strikers from the glass barrier between them and visitors from their relatives and inspections humiliation and punishment to the small.

14. RELEVANT: The transfer of two suffered minor injuries on Jerusalem hospital inside the camp and transport five others to al-hamshari Hospital of the organization of the Liberation of Palestine on the outskirts of the camps, as well as the knowledge of medical sources.

15. RELEVANT: It is noteworthy that there are approximately detained Palestinian out of eight in the prisons of the Israeli hunger strike as of August August, 2004 according to the club of the Palestinians in Bethlehem in the West Bank is the largest of the organizations defending prisoners of the Palestinians.

16. RELEVANT: The Muslim Brotherhood group in a statement today 8/23/2004, Monday 8/23/2004 all Muslims to fast Wednesday 8/18/2004 and Thursday 8/19/2004 in solidarity with the detainees Palestinians on hunger strike in prison .

17. NOT RELEVANT: He added: After tomorrow 8/17/2004 will join the rest of the prisoners in the general strike, the forces of Israeli today 8/16/2004 into the prisons strike and confiscated milk, salt, sugar and this demonstrates a negative intentions.

18. RELEVANT: Damascus 16-8 AFP - Scores of Palestinians on Monday 8/16/2004 a protest in front of the headquarters of the International Committee of the Red Cross in Damascus support Lamtaqlin Palestinians on hunger strike in Israel.

19. RELEVANT: And in solidarity with the prisoners on hunger strike, started at least 70 Palestinians representing different Palestinian national and Islamic factions and political forces Saturday 8/21/2004 strike open for food, said Arab member of the Knesset 48/218 Israeli parliament that the Palestinians inside Israel would announce general strike on Wednesday 8/18/2004 next day of solidarity with the prisoners of the Palestinians.

20. NOT RELEVANT: He said that the strike and collective hunger is aimed at spoiling the Israeli policy aimed at preventing prisoners Security of the planning of terrorist attacks in their cells, saying that the strike organized by the Islamic Resistance Movement, HAMAS and Islamic Jihad.

21. RELEVANT: And the flow of demonstrators took to the streets of Hebron, the meat and Tubas in the West Bank and Gaza city Hebron, carrying banners calling for the release of detainees and of releasing the prisoners of the brave. The demonstrators dozens of masked: some green books on such and reference to the Al-Aqsa Mosque.

22. NOT RELEVANT: It is noteworthy that Israel has more than 8 Palestinian prisoners for security reasons. Among the more than 90 women and 360 children, according to the International Committee of the Red Cross Organization of the United Nations Children.

23. RELEVANT: And around 1700 Palestinian prisoners out of eight thousand detainees in the prisons of the Israeli strike by one to improve prison conditions. The strike prisons whiff Wayshil in south Israel Wahdarim north of Tel Aviv.

24. NOT RELEVANT: Palestinians went on to say that the Israelis had brought the last night of the criminal Jews to prison Nafha and entered the sections prisoners the Palestinians and started the meat in a provocative step.

25. NOT RELEVANT: On the other hand, the confrontations light with stones in the center of the town of Hebron, during which the Palestinians threw stones from the army.

26. NOT RELEVANT: He called on the Palestinians strikers on Thursday 8/26/2004 organizations defending human rights, to intervene to avert a humanitarian disaster.

27. NOT RELEVANT: However, they remained in prison because the Palestinian leader Yasser Arafat.

28. NOT RELEVANT: The decision includes all prisoners of the strikers and 3200 detained in different prisons.

29. NOT RELEVANT: The saib negotiations minister in the Palestinian National Authority PNA on his readiness to meet with Israeli officials in order to follow up the file of POWs and detainees Palestinians on hunger strike for days.

30. NOT RELEVANT: Palestinian blamed the official Palestinian responsibility of the Government of Israel that could lead to repercussions.

31. NOT RELEVANT: The Palestinian President Yasser Arafat today 8/24/2004 Tuesday on the crimes committed against POWs of Palestinians in the prisons of the Israeli who thousands of their hunger strike for ten days.

B.2 Answers to GALE Year 1 Go/No-Go Evaluation Template 1 questions, Version 2 (with two-step IR)

GNG001 LIST FACTS ABOUT EVENT [Vice President Dick Cheney's shooting of Harry Whittington]
source date constraint: 11 February 2006 – 15 February 2006
source constraint: broadcast speech documents

1. NOT RELEVANT: his office and one witness suggested whittington was at fault because he failed to alert mr. cheney that he was rejoining the hunting party after searching for a down to earth.

2. NOT RELEVANT: vice president dick cheney under pressure goes public tonight about shooting his hunting companion.

3. NOT RELEVANT: when vice president dick cheney accidentally shot a member of his hunting party.

4. NOT RELEVANT: its different in the white house it 's different than the president he ought the president has a traveling team of reporters that follows him there is a certain communications flow that happens the vice president of often.

5. NOT RELEVANT: secondly uh oh you had two prominent democrats who are meeting with the president the vice president a small little is this morning here at the white house and came forward and said again using this as a metaphor.

6. NOT RELEVANT: vice president dick cheney 's been pretty mom since his shooting accident last weekend.

7. NOT RELEVANT: a half hour later he was in the white house watching the doctors televised press conference updating whittington condition that around one thirty the vice president called whittington wish him well.

8. NOT RELEVANT: but yes for a special edition of all the president 's men comes to d. v. d. today is a fabulous tribute to an important film

9. NOT RELEVANT: did the vice president follow hunting safety guidelines.

10. NOT RELEVANT: of the vice president away is an office handled robin first of all its becoming a distraction that is the primary concern of the white house the president is uh touting his health care plan and people keep asking at least a press corps pressing on some of these issues unanswered questions around the hunting accident.

GNG004 LIST FACTS ABOUT EVENT [The Hamas victory in Palestinian parliamentary elections]
source date constraint: 01 January 2006 – 01 May 2006

1. NOT RELEVANT: Israel, the United States and the European Union, which provide funding for a majority of the Palestinian Authority 's budget, have threatened to sever financial ties with the government if Hamas does not renounce violence and recognize Israel.

2. RELEVANT: Hamas landslide victory in Palestinian parliamentary elections last month prompted threats from the United States and European Union to cut off massive aid to the Palestinians unless the group recognizes Israel and renounces violence.

3. NOT RELEVANT: In a message issued to the people of Haiti on the eve of presidential and parliamentary elections, Annan said the elections offer an opportunity for Haiti to move away from violence and uncertainty towards a future of peace and stability.

4. NOT RELEVANT: The EU considers Hamas to be a terrorist group and has threatened to cut off millions of dollars in aid unless the group renounces violence and recognizes Israel.

5. NOT RELEVANT: Palestinian leader Mahmoud Abbas on Saturday asked Hamas to form the next Palestinian government, but demanded that the Islamic militants recognize existing peace deals and back his moderate policies, including negotiations with Israel, as the only strategic choice of the Palestinians.

6. RELEVANT: Exiled Hamas leaders from Syria joined Hamas leaders from Gaza in a series of meetings this week in Cairo to try to hammer out the movements plans for a new Palestinian government after last months landslide election win.

7. NOT RELEVANT: The militant group Hamas has decided to name Jamal al-Khudairi, a Gaza businessman who ran for parliament as an independent with Hamas backing, as its candidate for Palestinian prime minister, a top Hamas official said Wednesday.

8. RELEVANT: Regev made the statement shortly after Russian President Putin said at a press conference in Madrid that Russia was ready to invite Hamas members for talks in Moscow in the near future following the group 's landslide victory in last months Palestinian legislative elections.

9. RELEVANT: Hamas, holding the majority of parliament seats after winning the Jan. 25 parliamentary elections, is expected to form a new Palestinian government.

10. RELEVANT: Israel supports the Quartet decision, of which Russia was a party to, that there should be no political dialogue with Hamas until Hamas recognizes Israel, abandons violence and accepts the signed agreements, Israeli Foreign Ministry Spokesman Mark Regev was quoted by local newspaper The Jerusalem Post on its online edition as saying.

11. NOT RELEVANT: Many lawmakers have indicated they would oppose U.S. aid to a Hamas -led Palestinian government unless the group renounces violence.

12. NOT RELEVANT: The militant Palestinian group Hamas faces a historic choice at talks set for this month in Moscow whether to give up violence and recognize Israel, France 's prime minister said.

13. RELEVANT: Israel urged Russia on Thursday to follow the Quartet Committee 's consensus on the Palestinian Islamic Resistance Movement Hamas shortly after Russian President Vladimir Putin said he would invite Hamas for talks.

14. NOT RELEVANT: A new Palestinian parliament dominated by the militant group Hamas was installed here Saturday, and immediately, President Mahmoud Abbas and Hamas lawmakers

set out on a collision course over the need to honor existing agreements with Israel and conduct negotiations to achieve Palestinian statehood.

15. NOT RELEVANT: Russia has said it will press Hamas, which has killed hundreds of Israelis in suicide attacks, to abandon violence and recognize Israel 's right to exist.

16. NOT RELEVANT: Prime Minister Sali Berisha 's conservative government is planning to cut more than 2,000 jobs and reduce the number of ministries to 14 from 18, as part of a drive to reduce rampant government corruption in one of Europe 's poorest countries.

17. NOT RELEVANT: The militant Islamic Jihad group rejects the idea of a long-term truce with Israel and will not join a Hamas -led government, a leader of the group said Wednesday.

GNG005 LIST FACTS ABOUT EVENT [Appearance of human cases of bird flu in China]
source date constraint: 01 October 2005 – 01 May 2006

1. NOT RELEVANT: The virus has so far only been detected in birds passing through Hungary, and the aim is to prevent it from crossing over to local birds, said Gyurcsany, adding that should birds bred in Hungary be infected, vaccine would be available to protect people in close contact with them.

2. RELEVANT: Up to 15,000 fowl in Yijing, a town in China 's northern Shanxi province, were found dead late last week and tested positive for the H5N1 strain, Xinhua said.

3. NOT RELEVANT: Because migratory birds will return from Africa to Europe in the spring, there is a danger that wild birds could bring the poultry disease into the country, the statement said.

4. NOT RELEVANT: Officials fear the virus will spread in the spring when birds start to migrate, and Nabarros visit was intended as part of a U.N. effort to provide assistance to Ukraine.

5. RELEVANT: Hong Kong authorities on Thursday investigated two new suspected cases of bird flu, a dead chicken and a wild bird found near a local school, and the health secretary said he wont be surprised if more birds fall ill.

6. RELEVANT: Earlier Wednesday, China reported its 29th reported outbreak of the virus in birds since Oct. 19.

7. NOT RELEVANT: The two children suspected to have contacted the deadly H5N1 bird flu virus in the northern Nigerian state of Kaduna have been confirmed healthy and free from the virus, the official News Agency of Nigeria reported on Tuesday.

8. RELEVANT: Tamiflu is one of the drugs that could be used to help contain a possible human flu pandemic in the event that the bird flu virus mutates into a strain that can easily spread from human to human.

9. NOT RELEVANT: If the cases prove positive, it could mean the virus has spread from Kurdistan by infected migratory fowl that have passed it onto domestic bids or through Iraqis delivering infected birds from the north.

10. RELEVANT: Meanwhile, in Hong Kong on Wednesday, government health officials said a dead egret found in a suburb of Hong Kong has tested positive for H5N1 strain.

11. RELEVANT: Bird flu has ravaged Asian poultry stocks since late 2003, killing or forcing the slaughter of millions of birds.

12. NOT RELEVANT: While most of the human infections have been linked to direct contact with sick poultry, experts have warned that the virus could mutate into a form that is easily transmitted between people, sparking a global flu pandemic that could kill millions.

13. NOT RELEVANT: Azar, speaking to reporters, said the alliance would help prepare for any possible human flu pandemic and respond to those cases where bird flu has infected poultry and people.

14. RELEVANT: He said some patients could be living in environments that became contaminated with the virus through unknown channels, with the virus in those cases not necessarily causing large-scale deaths among animals.

GNG006 LIST FACTS ABOUT EVENT [Plots or attacks against US soldiers in Kuwait]
source date constraint: 01 January 2002 – 30 January 2006

1. RELEVANT: Kuwaiti security forces have detained up to eight soldiers suspected of planning to attack US forces in the emirate, the Arab Times reported Tuesday.

2. RELEVANT: Most attacks were carried out by fundamentalist Muslims who do not approve of the U.S. military presence in Kuwait.

3. NOT RELEVANT: United States Defense Secretary Donald Rumsfeld said America might need to send additional forces to Iraq where more well-organized attacks against US soldiers are emerging, the New York Times reported Monday.

4. NOT RELEVANT: Last weeks attacks in Riyadh, which left at least 34 dead and 194 injured, indicate that the al-Qaeda network remains active and highly capable of carrying out attacks, the official added on condition of anonymity.

5. NOT RELEVANT: The US embassy in Kuwait warned December 15 it had credible information that terrorist groups were preparing to carry out attacks in the emirate in the near future.

6. NOT RELEVANT: Germany plans to withdraw from Kuwait a unit specialized in detecting nuclear, biological and chemical warfare that was deployed as part of the U.S. -led war against terrorism, the defense minister said Wednesday.

7. NOT RELEVANT: Akbar is the only person charged in the grenade attack that killed two U.S. officers and wounded 14 other soldiers on March 23.

8. RELEVANT: The soldiers, some of them high-ranking officers, were detained and questioned a week ago over the plot, a Kuwaiti security source was quoted as saying.

9. RELEVANT: Around 25,000 US soldiers are stationed in Kuwait, which is also the main transit point for other coalition forces travelling to Iraq.

10. NOT RELEVANT: The unit remained in place during the war in neighboring Iraq despite the German government 's staunch opposition to U.S. military action and its refusal to contribute any troops.

11. NOT RELEVANT: The Kuwait mission fell under Operation Enduring Freedom, launched against the al-Qaida terror network following the Sept. 11, 2001 attacks in the United States.

12. NOT RELEVANT: On late Monday, Kuwaiti army spokesman Brigadier Yussef Al-Mulla told the Kuwait News Agency that the military intelligence service is questioning some soldiers, following information concerning their intention to carry out an attack on friendly forces, but he did not specify who the friendly forces were.

13. NOT RELEVANT: Kuwait, liberated from Iraqi occupation by a US -led coalition in the 1991 Gulf War, has witnessed three serious shooting incidents, two fatal, involving Americans since October 2002.

14. NOT RELEVANT: The US soldiers vehicle ran over a mine, and the blast killed four soldiers, she confirmed through telephone, adding that the incident happened when US troops are on a joint routine patrol mission with Afghan National Army in Logar Province.

15. NOT RELEVANT: Staunch US ally Kuwait last week raised its state of alert almost to the maximum, boosting security around the country in the biggest show of force since the March 2003 launch of the US -led war in Iraq.

16. NOT RELEVANT: Last month, Germany withdrew 140 soldiers who reinforced the contingent during the war.

17. NOT RELEVANT: The US intelligence community assesses that attacks against US and Western targets overseas are likely, the official said, noting that the possibility of attacks on US soil could not be ruled out.

18. RELEVANT: A Kuwaiti army spokesman said Thursday that two Kuwaiti soldiers were to stand trial on suspicion of plotting attacks on US and other foreign forces in the country, which served as the main launchpad for the war.

19. NOT RELEVANT: The US embassy in Kuwait earlier Monday warned its citizens that assailants in a black car driving around Kuwait planned to randomly attack Westerners, in a message posted on the embassy website.

20. NOT RELEVANT: KUWAIT CITY, Jan 10 AFP - Kuwaiti security forces fatally wounded a wanted militant Monday in a deadly shootout in which two policemen were killed amid a US security alert, the interior ministry said.

21. NOT RELEVANT: McKiernan said the Army 's 3rd Infantry Division, which had been planning to return to the United States in June, was going to remain in Iraq until commanders decided they were no longer needed.

22. NOT RELEVANT: Two Kuwaiti security men were killed and two others wounded in a shootout with gunmen in a suburb south of the capital on Monday, Kuwait television reported, quoting a security source.

23. NOT RELEVANT: Earlier on Monday, Kuwait began to ease tight security measures introduced almost two weeks ago when the emirate raised its state of alert almost to the maximum in the biggest show of force since the March 2003 launch of the US-led war in Iraq.

24. NOT RELEVANT: Three soldiers with the Army 's 101st Airborne Division were killed in an ambush here early Thursday morning that involved what one local resident said was an attack by insurgents armed with rocket-propelled grenades and Kalashnikov assault rifles.

25. NOT RELEVANT: The arrests follow Kuwait 's stepping up its internal security in recent days.

26. NOT RELEVANT: An American soldier was killed and seven other soldiers were injured in the attack in falujah 50 km west of Baghdad, according to the leadership of the Central America.

27. NOT RELEVANT: But Germany said the unit would also have been ready to help if Iraq attacked Kuwait or U.S. forces there with weapons of mass destruction.

28. NOT RELEVANT: The Kuwait army said Monday it has arrested a number of soldiers suspected of planning to attack US forces in the emirate, an army spokesman said.

29. NOT RELEVANT: This was a combined joint task force the United States, the Czech Republic and Germany posted defense forces against nuclear, biological and chemical weapons to Kuwait, Defense Minister Peter Struck told ZDF television.

30. NOT RELEVANT: U.S. authorities will not disclose the exact figure of American forces in Kuwait, which was used by coalition forces to invade neighboring Iraq in March 2003.

31. NOT RELEVANT: The military intelligence security service is questioning some soldiers, following information concerning their intention to carry out an attack on friendly forces, said army spokesman Yussef al-Mulla, the Kuna agency reported.

32. NOT RELEVANT: Around 25,000 US soldiers are stationed in Kuwait, which is used as a transit point for US and other coalition troops headed for Iraq.

33. NOT RELEVANT: The militant, named as Fawaz Talaiq al-Oteibi, died in hospital after the shootout which erupted as security men came to arrest him in a Kuwait City suburb, said an interior ministry statement.

34. NOT RELEVANT: A security source announced that two security men were killed and two others wounded in a shootout with gunmen in Hawally, state-run TV announced, without providing further details.

35. RELEVANT: The Kuwaiti army said Monday it had arrested a number of soldiers who were planning to attack friendly forces in the emirate, two weeks after the United States warned of the increased possibility of militant attacks.

36. NOT RELEVANT: He said that for the first time the attacks against American troops were being coordinated regionally and possibly nationally by remnants of ousted Iraqi leader Saddam Hussein 's security forces, guerrilla fighters and Iraqi prisoners released before the war.

37. NOT RELEVANT: On Thursday, a U.S. soldier was killed when his convoy came under fire from a rocket -propelled grenade on a supply route through Iraq, bringing to nine the number of American soldiers who have died around the country this week.

38. NOT RELEVANT: Kuwait is holding three members of its armed forces who are suspected of plotting to attack allied forces, the defense minister said Tuesday.

GNG025 LIST FACTS ABOUT EVENT [The shut down of the Cernavoda nuclear power plant]

location constraint: Romania

source date constraint: 24 August – 01 December 2003

1. NOT RELEVANT: Romania 's Jewish community protested the remarks, and the government eventually acknowledged that Romania 's wartime leaders deported and killed Jews.

2. RELEVANT: Its the first time the plant in Cernavoda, some 200 kilometers 125 miles east of Bucharest, has encountered water level problems since it opened seven years ago.

3. RELEVANT: The plant was turned off on Aug. 24, after water levels dropped as low as 1.5 meters nearly 5 feet down from its usual level of almost seven meters 23 feet.

4. RELEVANT: The water level in the Danube River at Cernavoda village, where the reactor is located, fell to a depth of less than three meters 10 feet on Saturday, down from its usual level of almost seven meters 23 feet.

5. RELEVANT: A nuclear power plant was shut down Sunday because a record drought left insufficient water to cool down the reactor.

6. NOT RELEVANT: The manual, which will be given to schools around Romania, was issued three months after the Romanian government said in a statement that there had been no Holocaust inside Romania 's borders.

7. NOT RELEVANT: Romania 's traditionally good ties to Israel were strained in June after the government claimed there was no Holocaust inside Romania 's borders.

GNG041 LIST FACTS ABOUT EVENT [a ferry crash] (location constraint: Canada)

location constraint: Canada

source date constraint: 01 February 2005 – 30 June 2005

source constraint: broadcast speech documents

1. NOT RELEVANT: Canada backed away from a pledge to join the US Ballistic Missile Defense BMD system, US ambassador to Canada Paul Cellucci said in an interview aired Sunday.

2. NOT RELEVANT: Martin had promised a new era of Canada - U.S. relations after bitter divisions over the war in Iraq, and Americans have warned relations would deteriorate further if Canada refused to join the missile plan.

3. NOT RELEVANT: In the defence package, three billion dollars 2.4 billion US will pay for the increase of 5,000 regular troops and 3,000 reservists already announced by the government ; 3.2 billion dollars 2.6 billion US to improve training of Canada 's forces, repairing infrastructure, and stepping up the provision of supplies and repairs which have been the subject of complaint by the military for several years.

4. NOT RELEVANT: US Secretary of State Condoleezza Rice announced Tuesday she was postponing a visit to Canada after conveying her disappointment over Canada 's withdrawal.

5. NOT RELEVANT: Statistics Canada also said that in December finished-product inventories rose to a record 21.6 billion Canadian dollars 17 billion US dollars, surpassing the previous high of 21.4 billion hit in June, 2001.

6. NOT RELEVANT: Canada 's trade surplus has fallen to about 3.2 billion US dollars in January from about 4.1 billion US dollars in December, Statistics Canada said Friday.

7. NOT RELEVANT: Domestically, the budget provides 222 million dollars 179 million US for increased security on the Great Lakes and St. Lawrence Seaway, which border the United States, and for increased police presence in Canada 's ports.

8. NOT RELEVANT: Natural gas exports fell 16.7 percent to 2.5 billion Canadian dollars about 2 billion US dollars, accounting for more than half of the drop, while crude oil exports fell by 11.3 percent to 2.2 billion Canadian dollars about 1.8 billion US dollars.

9. NOT RELEVANT: We simply cannot understand why Canada would in effect give up its sovereignty its seat at the table to decide what to do about a missile that might be coming

towards Canada, the outgoing ambassador, who had vigorously urged Canada to sign on the plan during his tenure, told reporters in Ottawa immediately after Martin 's announcement.

10. NOT RELEVANT: Canada announced its decision on the missile defense system last week, setting off a prickly exchange between the U.S. ambassador to Canada and Canadian Prime Minister Paul Martin.

11. NOT RELEVANT: Canada said Wednesday it would spend an extra 38 million Canadian dollars 31 million US this year in the global fight against tuberculosis.

12. NOT RELEVANT: And that could mean redressing the almost 90 countries that Canada sends an annual amount of less than 5 million Canadian dollars 4 million US dollars.

13. NOT RELEVANT: Overall, Canadian exports fell 1.6 percent from December to 35. 9 billion Canadian dollars about 29.7 billion US dollars, while imports rose 1.9 percent to 31.9 billion Canadian dollars about 26.5 billion US dollars.

14. NOT RELEVANT: Canada 's Liberal Prime Minister Paul Martin leads a minority government and his party 's lawmakers had lobbied him not to take part in the program, which is highly unpopular in Canada, particularly in Quebec.

15. NOT RELEVANT: Why run this risk before we can be confident that Canada is enforcing its own regulations? said Conrad, charging that Canada 's regulatory measures for cattle and beef were insufficient.

16. NOT RELEVANT: Under the plan, Canada would contribute approximately 172 million Canadian dollars about 140 million US dollars over the next five years to the International Development Association of the World Bank and to the African Development Fund.

GNG044 LIST FACTS ABOUT EVENT [hunger strikes by Palestinians in Israeli jails]
source date constraint: 01 August 2004 – 31 August 2004

1. NOT RELEVANT: The 800 Palestinian prisoners in prison today, Friday, and even the two their hunger strike, which began mid-August, half of the prisoners the Palestinians nearly eight thousand detained in prisons, also announced the club of the prisoner.

2. NOT RELEVANT: RAMALLAH West Bank 30-8 AFP - The Club of the Palestinian Monday that approximately 800 detained in prison the Israeli resumed their hunger strike after the Adarralsjun Israeli earlier promises made by them.

3. RELEVANT: Arafat at the beginning of the Llasra on hunger strike in the prisons of the Israeli announced that it had been agreed with the Palestinian leadership and the forces that today is a day of fasting in solidarity with the prisoners.

4. NOT RELEVANT: Arafat said after meeting with the Canadian representative to the Palestinian Authority Steve Haybard entered the strike of the Palestinian prisoners Day of his tenth the crimes perpetrated against our prisoners Wasiratna cannot be ignored.

5. RELEVANT: Source said Arab official here today that the Council of the League of Arab States held here today held an emergency meeting at the level of permanent delegates to discuss the grave situation and the deterioration of the prisoners and detainees the Palestinians and the Arabs in the prisons.

6. RELEVANT: And some 7500 prisoners in Israeli prisons since the beginning of a hunger strike of the Israeli the achievement of a series of demands related to their living conditions inside the prisons they say humiliation and violation of all religions and laws.

7. RELEVANT: It demands that the prisoners, in accordance with the club of the prisoner, to stop the policy of inspection of Walmhal, end the policy of fines and end the policy of repression and aggression on the prisoners, end the policy of storming the Chambers, and visits to the prisoners of war and the removal of the glass rooms for the visits Laysmh to sixty of them.

8. RELEVANT: The statement pointed to the conditions of the families in the prisons of the Israeli that made the lives of the prisoners of death Watwaziyah and paid by violations of them to be coexistence with him and kept silent about the and the prisoners to fight the bowel, for the sake of dignity and decent life.

9. RELEVANT: Hisham Abdelrazek Affairs Minister prisoners of war and the editors of the Palestinians in the wake of the meeting of the government of the Palestinian told reporters that the government has decided today to the United Nations to the claim of prisoners Palestinian prisoners of war.

10. NOT RELEVANT: The Council of the League of Arab States at the level of permanent delegates held an emergency meeting yesterday in which he called for holding an urgent meeting Watara International Commission for Human Rights to consider the tragic situation which have suffered by the Palestinian prisoners of the Arabs in prisons.

11. NOT RELEVANT: Pursuant to a general strike in the city of Hebron in the West Bank today

at the invitation of the Palestinian national forces in solidarity with the strike of prisoners Palestinians on hunger strike launched by the prisoners.

12. NOT RELEVANT: The Radio Israel quoted an as saying following a meeting of Ladirab prisoners that Israel would not accept any of the demands of the prisoners.

13. NOT RELEVANT: He said that the strike and collective hunger is aimed at spoiling the Israeli policy aimed at preventing prisoners Security of the planning of terrorist attacks in their cells, saying that the strike organized by the Islamic Resistance Movement, HAMAS and Islamic Jihad.

14. RELEVANT: The Jordanian news agency that the president of the Federation of Women the Jordanian safe Alza bi revealed during their meeting, Millah at the Headquarters of the Red Cross to the demands of the Palestinian prisoners and the Jordanian fair and must be the response and the international pressure to respond to the international conventions and Human Rights expressed its concern about the Israel in its policy towards the Wamilha that there would be an international response to their demands focused on the issue of the prisoners, without charge or trial.

15. RELEVANT: Today, Tuesday club of Palestinian captive of Pope John Paul the second to intervene on the side of the detainees Palestinians on hunger strike in Israel.

16. RELEVANT: More than in the person of today in the West Bank to express their support for the detainees Palestinians who carry on a hunger strike in Israeli jails.

17. RELEVANT: And the flow of demonstrators took to the streets of Hebron, the meat and Tubas in the West Bank and Gaza city, carrying banners calling for the release of detainees and of releasing the prisoners of the brave.

18. RELEVANT: He called out to these committees to take the necessary measures for authorities in an attempt Bilatfaqiat and relevant including the application of the Geneva Conventions on the prisoners of the Palestinians, according to which dealing with them as prisoners of war and the crimes and violations.

19. RELEVANT: The 1700 Palestinian prisoners out of eight thousand detainees in the prisons of the Israeli went on hunger strike yesterday Sunday to improve prison conditions.

20. NOT RELEVANT: He said Alnadi in a statement that a number of prisoners in prisons and al-khayam Ramla Watlmund Wajlbu and the Negev desert of prisoners and detainees in isolation

Alanfaradi in Sajn Il- rmalh and other children in Telmond joined b 1700 prisoners in prisons Shatta and Nafha Walsb Wahdarim who entered their hunger strike on his fourth.

21. NOT RELEVANT: He added Qaraq, president of the club, told Agence France Presse that the detainees who decided to return the strike as Adarhasjn allowing them to contact with the rest of the prisoners as previously promised.

22. NOT RELEVANT: He added, we must wait for minutes before talk about resuming the strike, pointing out that between and detained were concerned Balidrab in Ashkelon and in Israeli prisons and not four thousand according to the Palestinians.

23. RELEVANT: And half of 8000 detained in Israeli prisons in the movement of hunger strike, which started the one to improve the conditions of their detention.

24. NOT RELEVANT: He said the club is that the strike will start between 15 and 18 of this month will be announced at the beginning of each of the prisons of Beersheba and Nafha Wahdarim, Mamour, the strike that they are the other prisons after three days.

25. NOT RELEVANT: The Israeli Laflar spokesman of the Prisons Department told Agence France Presse that was involved in the strike in civilian prison, affirming that the strike did not extend to military detention.

26. NOT RELEVANT: The Palestinian officials today Saturday that about eight thousand Palestinians will begin a hunger strike in order to improve the conditions of their detention while the Israeli authorities in advance of any compromise.

27. RELEVANT: It is noteworthy that three to four thousand Palestinian out of a total of 7500 in the prisons of the Israeli The fifteenth of August on hunger strike demanding an improvement in their conditions of detention.

28. NOT RELEVANT: The decision includes all prisoners of the strikers and 3200 detained in different prisons.

29. RELEVANT: The bodies in a statement made by the during the sit-in to the President of the Assembly of the International Red Cross in Jordan, Millah all international and humanitarian organizations and the International Committee of the Red Cross of pressure on Israel to respect and implement the fourth Geneva Convention and international humanitarian law with regard to the prisoners of war in Israeli prisons and the pressure for the release of prisoners are all in their release Alasirat and children under the age.

30. NOT RELEVANT: The channel Al-Jazeera satellite channel that the suspension of the strike came to the prison authorities to part of the demands made by the prisoners at the beginning of their strike 0

31. RELEVANT: The statement pointed out that 3500 other detention centres of the Israeli army Badawa today a solidarity with the prisoners strikers in the central prisons include the strike An food today for one day and the start tomorrow in the province of prisons and the return to the hunger strike Friday one day.

32. NOT RELEVANT: The Palestinian President Yasser Arafat today Tuesday on the crimes committed against POWs of Palestinians in the prisons of the Israeli who thousands of their hunger strike for ten days.

33. NOT RELEVANT: The prisoners in the prisons of the Israeli, and their numbers in the thousands, their determination to implement the open strike on hunger strike if demands were not met in rejecting the Israeli authorities that Hamas and Jihad to strike.

34. NOT RELEVANT: The Director of the Information Department of the Arab League Mahmoud Abdel Aziz, in a press statement following the meeting that the meeting held at the request of Palestine to mobilize public opinion against the Israeli practices in the prisons and the movement of international solidarity in this issue with a view to put pressure on Israel to abide by the rules of international law adequate in this regard and formulating a plan Arab action on the international scene regarding this issue.

35. RELEVANT: And in solidarity with the prisoners on hunger strike, started at least 70 Palestinians representing different Palestinian national and Islamic factions and political forces Saturday strike open for food, said Arab member of the Knesset 48/218 Israeli parliament that the Palestinians inside Israel would announce general strike on Wednesday next day of solidarity with the prisoners of the Palestinians.

36. NOT RELEVANT: The; prison al-khayam announced the end of last week their approval for the suspension of the strike which they mid-August after approved the prison administration to enter into negotiations with them on their demands and allowing them to contact with their strike in the other prisons.

37. RELEVANT: Systems thousands of Palestinians massive demonstrations today in the cities of the West Bank and Gaza Strip, to express their support for the detainees Palestinians on hunger strike in Israeli jails.

38. NOT RELEVANT: The supreme body to follow up on the POWs and detainees a statement that called for all the people of the Palestinians at home and abroad and occupied Palestine in and its friends in the world to initiate the open strike on hunger strike as of today.

39. RELEVANT: He also Moussa condemnation and dissatisfaction of the statements of the minister of internal security the Hanghbi in which he referred to its disregard for the lives of the prisoners and detainees Palestinians who started their hunger strike in the prisons and detention camps.

40. RELEVANT: It is noteworthy that about eight thousand Palestinian detainees in the prisons of the Israeli went on hunger strike last Sunday in an attempt to improve the conditions of Itiqalahm, also of the decision to join the prisoners Alakharun as of tomorrow.

41. NOT RELEVANT: He went on to say Qaraq and we fear that the prisons department to try to break the strike to attempt to force the detainees eating the force, as was the case in strike whiff in 1980 which led to the death of two detainees.

B.3 Answers to GALE Year 1 Go/No-Go Evaluation Template 1 questions, Version 3 (with two-step IR and information packageing procedure)

GNG001 LIST FACTS ABOUT EVENT [Vice President Dick Cheney's shooting of Harry Whittington]
source date constraint: 11 February 2006 – 15 February 2006
source constraint: broadcast speech documents

1. NOT RELEVANT: vice president dick cheney is back in washington after a weekend hunting accident.

2. NOT RELEVANT: the vice president dick cheney accidentally shot a man during a quail hunt at a political supporters ranch making seventy year old harry whittington.

3. NOT RELEVANT: hunting buddies have something to say about vice president dick cheney safety record with a rifle next why they think the criticism after saturdays shooting incident is not a real picture.

4. NOT RELEVANT: another serious question tonight of course did the vice president follow hunting safety standards.

5. NOT RELEVANT: his office and one witness suggested whittington was at fault because he failed to alert mr. cheney that he was rejoining the hunting party after searching for a down to earth.

6. NOT RELEVANT: when vice president dick cheney accidentally shot a member of his hunting party.

7. NOT RELEVANT: word his victim harry whittington suffered a mild heart attack produced the first official written statement from cheney 's office and knowledge and that the incident had even occurred.

8. NOT RELEVANT: its different in the white house it 's different than the president he ought the president has a traveling team of reporters that follows him there is a certain communications flow that happens the vice president of often.

9. NOT RELEVANT: david letterman having a laugh and vice president dick cheney 's expense obviously.

10. RELEVANT: the man the vice president dick cheney shot during a hunting trip is recovering from a minor heart attack.

11. NOT RELEVANT: he was accidently shot with palates by vice president dick cheney during a quail hunting trip at a ranch in texas.

12. NOT RELEVANT: and the white house have faulted whittington for failing to alert mr. cheney that he was rejoining the hunting line.

13. NOT RELEVANT: the vice president is back in washington after a weekend hunting accident he shot and injured a member of his hunting party saturday while they were shooting quail at a friend 's ranch in texas.

14. NOT RELEVANT: the storm of anger scathing criticism today from members of both parties for homeland security secretary michael chertoff and his response to hurricane katrina.

15. NOT RELEVANT: vice president dick cheney under pressure goes public tonight about shooting his hunting companion.

16. NOT RELEVANT: a national failure of leadership to an initiative that is what a no holds barred congressional report says about the government 's response to hurricane katrina.

17. NOT RELEVANT: vice president dick cheney 's been pretty mom since his shooting accident last weekend.

18. NOT RELEVANT: of the vice president away is an office handled robin first of all its becoming a distraction that is the primary concern of the white house the president is uh touting his health care plan and people keep asking at least a press corps pressing on some of these issues unanswered questions around the hunting accident.

19. NOT RELEVANT: since vice president dick cheney accidentally shot his hunting companion saturday evening.

20. NOT RELEVANT: suzanne malveaux has the latest now on the white house 's attempts to quiet those critics.

21. NOT RELEVANT: the accident itself and what has this honey accident done in terms of illuminating.

22. NOT RELEVANT: and heres a look at whats going on the man the vice president dick cheney shot during a hunting trip is recovering from a minor heart attack.

23. NOT RELEVANT: did the vice president follow hunting safety guidelines.

24. NOT RELEVANT: heres what youll see tonight when you turn to ten beginning at seven thirty dont miss the olympic zone for source for local olympic coverage at eight oclock a. b. c. olympic coverage live from torino then be sure to stay with us for late edition of n. b. c. ten news at eleven.

GNG004 LIST FACTS ABOUT EVENT [The Hamas victory in Palestinian parliamentary elections]

source date constraint: 01 January 2006 – 01 May 2006

1. RELEVANT: Regev made the statement shortly after Russian President Putin said at a press conference in Madrid that Russia was ready to invite Hamas members for talks in Moscow in the near future following the group 's landslide victory in last month 's Palestinian legislative elections.

2. NOT RELEVANT: With a solid majority of parliament seats, Hamas is expected to dominate the next Palestinian government and could have a say over key security matters.

3. NOT RELEVANT: Following Hamas 's electoral triumph, the Quartet for Middle East peace – the United Nations, the United States, Russia and the European Union – demanded that Hamas give up violence and recognize Israel.

4. NOT RELEVANT: He accused Israel of trying to provoke Hamas when the world has been urging it to renounce violence.

5. NOT RELEVANT: A new Palestinian parliament dominated by the militant group Hamas was installed here Saturday, and immediately, President Mahmoud Abbas and Hamas lawmakers set out on a collision course over the need to honor existing agreements with Israel and conduct negotiations to achieve Palestinian statehood.

6. NOT RELEVANT: Russia has said it will press Hamas, which has killed hundreds of Israelis in suicide attacks, to abandon violence and recognize Israel 's right to exist.

7. NOT RELEVANT: The United States and the European Union, both of which regard Hamas as a terrorist organization, have threatened to cut off aid unless it gives up violence and recognizes Israel.

8. NOT RELEVANT: The EU considers Hamas to be a terrorist group and has threatened to cut off millions of dollars in aid unless the group renounces violence and recognizes Israel.

9. NOT RELEVANT: In a message issued to the people of Haiti on the eve of presidential and parliamentary elections, Annan said the elections offer an opportunity for Haiti to move away from violence and uncertainty towards a future of peace and stability.

10. RELEVANT: Hamas landslide victory in Palestinian parliamentary elections last month prompted threats from the United States and European Union to cut off massive aid to the Palestinians unless the group recognizes Israel and renounces violence.

11. RELEVANT: Hamas, holding the majority of parliament seats after winning the Jan. 25 parliamentary elections, is expected to form a new Palestinian government.

12. NOT RELEVANT: The voter picks one, puts it in the envelope and deposits the envelope in the ballot box.

13. NOT RELEVANT: Annan also expressed his confidence that the elections will prove a significant step in the work to build a more stable Haiti, and called on all parties to respect the outcome of the elections, and on the incoming leadership to demonstrate commitment to reconciliation and inclusiveness.

14. NOT RELEVANT: Israel, the United States and the European Union, which provide funding for a majority of the Palestinian Authority 's budget, have threatened to sever financial ties with the government if Hamas does not renounce violence and recognize Israel.

15. RELEVANT: Russian officials promised they would demand that Hamas recognize the state of Israel and abandon the use of violence.

16. NOT RELEVANT: Palestinian leader Mahmoud Abbas on Saturday asked Hamas to form the next Palestinian government, but demanded that the Islamic militants recognize existing peace deals and back his moderate policies, including negotiations with Israel, as the only strategic choice of the Palestinians.

17. RELEVANT: Exiled Hamas leaders from Syria joined Hamas leaders from Gaza in a series of meetings this week in Cairo to try to hammer out the movements plans for a new Palestinian government after last month 's landslide election win.

18. RELEVANT: Israel supports the Quartet decision, of which Russia was a party to, that there should be no political dialogue with Hamas until Hamas recognizes Israel, abandons violence and accepts the signed agreements, Israeli Foreign Ministry Spokesman Mark Regev was quoted by local newspaper The Jerusalem Post on its online edition as saying.

19. NOT RELEVANT: In an interview with the British Broadcasting Corp. in Cairo, Meshaal also said that Hamas would not renounce violence as it is entitled to resist what it regards as Israel 's occupation of Palestinian land.

20. NOT RELEVANT: The militant Palestinian group Hamas faces a historic choice at talks set for this month in Moscow whether to give up violence and recognize Israel, France 's prime minister said.

21. RELEVANT: However, the United States has threatened to cut its direct aid to the Palestinian government if Hamas, regarded by the United States as a terrorist group, fails to renounce violence against Israel and dismantle its armed wing.

22. RELEVANT: Israel urged Russia on Thursday to follow the Quartet Committee 's consensus on the Palestinian Islamic Resistance Movement Hamas shortly after Russian President Vladimir Putin said he would invite Hamas for talks.

23. NOT RELEVANT: Secretary of State Condoleezza Rice called on Haiti 's citizens and political parties to respect the results of presidential and parliamentary elections when they are announced.

24. NOT RELEVANT: UN chief Kofi Annan on Thursday appealed to the militant Palestinian Islamist group Hamas to transform itself into a political party following its shock election victory last month.

25. NOT RELEVANT: Many lawmakers have indicated they would oppose U.S. aid to a Hamas -led Palestinian government unless the group renounces violence.

26. NOT RELEVANT: The militant group Hamas has decided to name Jamal al-Khudairi, a Gaza businessman who ran for parliament as an independent with Hamas backing, as its candidate for Palestinian prime minister, a top Hamas official said Wednesday.

27. RELEVANT: The Quartet Committee urged Hamas, which is sworn to Israel 's destruction, to renounce violence, recognize the existence of Israel and abide by previous Palestinian accords with Israel at a London meeting on Jan. 30.

28. RELEVANT: Five days after Hamas won the Palestinian parliamentary elections, European ministers met on Jan. 30 in Brussels to urge the radical movement to renounce violence, recognize Israel and disarm.

29. NOT RELEVANT: The poll by the Geocartographia survey firm showed the new combination willing nine seats, down from 13 as separate parties in the current parliament.

30. NOT RELEVANT: A poll broadcast on Israel Radio on Thursday showed Kadima winning 38 seats, Labor 17 and Likud 15 in the 120-seat parliament.

GNG005 LIST FACTS ABOUT EVENT [Appearance of human cases of bird flu in China]
source date constraint: 01 October 2005 – 01 May 2006

1. NOT RELEVANT: While most of the human infections have been linked to direct contact with sick poultry, experts have warned that the virus could mutate into a form that is easily transmitted between people, sparking a global flu pandemic that could kill millions.

2. NOT RELEVANT: Almost all of the human deaths have been linked to contact with infected poultry, but experts fear the H5N1 virus could mutate into a form that spreads easily among people, possibly sparking a human flu pandemic.

3. NOT RELEVANT: The virus has so far only been detected in birds passing through Hungary, and the aim is to prevent it from crossing over to local birds, said Gyurcsany, adding that should birds bred in Hungary be infected, vaccine would be available to protect people in close contact with them.

4. RELEVANT: The Chinese woman 's infection was in coastal Fujian province and announced Wednesday, the same day that authorities reported Africa 's first cases of the deadly H5N1 strain, an outbreak on a poulty farm in Nigeria where no human infections were reported but 40,000 birds died.

5. RELEVANT: He said some patients could be living in environments that became contaminated with the virus through unknown channels, with the virus in those cases not necessarily causing large-scale deaths among animals.

6. NOT RELEVANT: The two children suspected to have contacted the deadly H5N1 bird flu virus in the northern Nigerian state of Kaduna have been confirmed healthy and free from the virus, the official News Agency of Nigeria reported on Tuesday.

7. RELEVANT: Earlier Wednesday, China reported its 29th reported outbreak of the virus in birds since Oct. 19.

8. RELEVANT: Tamiflu is one of the drugs that could be used to help contain a possible human flu pandemic in the event that the bird flu virus mutates into a strain that can easily spread from human to human.

9. RELEVANT: Hong Kong authorities on Thursday investigated two new suspected cases of bird flu, a dead chicken and a wild bird found near a local school, and the health secretary said he wont be surprised if more birds fall ill.

10. RELEVANT: However Mao reassured there was no evidence of the virus mutating into a strain that could be transmitted easily among humans, circumstances that health experts fear would cause a global pandemic that would kill millions of people.

11. RELEVANT: Bird flu has ravaged Asian poultry stocks since late 2003, killing or forcing the slaughter of millions of birds.

12. NOT RELEVANT: Meanwhile, Christina Carlson, a U.N. specialist working in the former Soviet Union, said that experts had studied the capacities of hospitals in Crimea to deal with cases of human infection and found they have sufficient capacity to handle such cases if they will come up.

13. NOT RELEVANT: If the cases prove positive, it could mean the virus has spread from Kurdistan by infected migratory fowl that have passed it onto domestic bids or through Iraqis delivering infected birds from the north.

14. NOT RELEVANT: Azar, speaking to reporters, said the alliance would help prepare for any possible human flu pandemic and respond to those cases where bird flu has infected poultry and people.

15. RELEVANT: Up to 15,000 fowl in Yijing, a town in China 's northern Shanxi province, were found dead late last week and tested positive for the H5N1 strain, Xinhua said.

16. RELEVANT: In four of these cases the health ministry later found the patients had close contacts with sick birds, although the agricultural ministry could still not determine a bird flu outbreak.

17. NOT RELEVANT: A further two wild ducks have tested positive for the H5N1 strain of bird flu virus in Italy, after eight swans had been confirmed by Italian authorities last week of carrying the deadly virus.

18. NOT RELEVANT: Officials fear the virus will spread in the spring when birds start to migrate, and Nabarros visit was intended as part of a U.N. effort to provide assistance to Ukraine.

19. NOT RELEVANT: China 's Health Ministry confirmed that the woman in Fujian, who was surnamed Liu, tested posited for the H5N1 strain, the official Xinhua News Agency said.

20. NOT RELEVANT: Its last big outbreak was in 1997, when bird flu killed six people, forcing the government to slaughter the territory 's entire chicken population of 1.5 million birds.

21. RELEVANT: China says that a 26-year-old woman in a southeastern province has contracted bird flu in the country 's 11th human case of the deadly virus, which has spread through birds from Asia to Europe and, most recently, Africa.

22. RELEVANT: Meanwhile, in Hong Kong on Wednesday, government health officials said a dead egret found in a suburb of Hong Kong has tested positive for H5N1 strain.

23. NOT RELEVANT: Officials fear the virus will spread in the spring when birds start to migrate.

24. NOT RELEVANT: Magnusson said those major outbreaks differed significantly in scope from the sporadic outbreaks in migratory birds found in other European countries such as Italy and Greece.

GNG006 LIST FACTS ABOUT EVENT [Plots or attacks against US soldiers in Kuwait]
source date constraint: 01 January 2002 – 30 January 2006

1. RELEVANT: Two Kuwaiti soldiers are to stand trial on suspicion of plotting attacks on US and other foreign forces, army spokesman Brigadier Yussef al-Mulla said Thursday.

2. RELEVANT: Around 25,000 US soldiers are stationed in Kuwait, which is also the main transit point for other coalition forces traveling to Iraq.

3. NOT RELEVANT: The suspect fired at police and then fled to a car where a group of his friends were waiting...

4. NOT RELEVANT: Kuwait on Friday raised its state of alert almost to the maximum, boosting security around the country in the biggest show of force since the US -led Iraq war in March 2003.

5. NOT RELEVANT: KUWAIT CITY, Jan 10 AFP - Kuwaiti security forces fatally wounded a wanted militant Monday in a deadly shootout in which two policemen were killed amid a US security alert, the interior ministry said.

6. RELEVANT: Al-Qaeda -linked militants who fought four bloody gunbattles with Kuwaiti security forces over the past month plotted to kidnap and execute US soldiers and Westerners, a newspaper reported Saturday.

7. NOT RELEVANT: The militant, named as Fawaz Talaiq al-Oteibi, died in hospital after the shootout which erupted as security men came to arrest him in a Kuwait City suburb, said an interior ministry statement.

8. NOT RELEVANT: One of the men in the car shot at the police killing two and wounding two others, the interior ministry said.

9. NOT RELEVANT: Kuwait meanwhile remained on a state of alert in the wake of violence that left 14 people dead as security forces continued the hunt for a number of fugitives after killing eight of them and capturing 14 others.

10. RELEVANT: A Kuwaiti army spokesman said Thursday that two Kuwaiti soldiers were to stand trial on suspicion of plotting attacks on US and other foreign forces in the country, which served as the main launchpad for the war.

11. NOT RELEVANT: Two other policemen were wounded in the firefight which came as the US embassy warned its nationals that militants at large in a car could randomly attack Westerners in the emirate.

12. RELEVANT: Around 25,000 US soldiers are stationed in Kuwait, which is also the main transit point for coalition forces travelling to and from Iraq.

13. NOT RELEVANT: Some 12,000 American civilians live in Kuwait alongside around 9,000 Europeans and some 1,000 Australians.

14. NOT RELEVANT: The US embassy warned on December 15 it had credible information that terrorist groups were preparing to carry out attacks in the emirate in the near future.

15. NOT RELEVANT: There are some 25,000 US troops stationed in staunch Washington ally Kuwait which serves as a transit point for coalition forces moving in and out of Iraq.

16. NOT RELEVANT: On late Monday, Kuwaiti army spokesman Brigadier Yussef Al-Mulla told the Kuwait News Agency that the military intelligence service is questioning some soldiers, following information concerning their intention to carry out an attack on friendly forces, but he did not specify who the friendly forces were.

17. NOT RELEVANT: Plainclothes police tried to arrest Oteibi as he went to return a rented car in Hawalli, some 10 kilometers six miles south of Kuwait City.

18. NOT RELEVANT: Hundreds of Palestinian security forces were being redeployed to the northern Gaza Strip Thursday with orders to prevent attacks by militants, after an agreement with Israel, security sources said.

19. NOT RELEVANT: The embassy is issuing this urgent message because it has received credible information that an individual or individuals moving about Kuwait in a black coloured small sedan intend to randomly attack Westerners, the message said.

20. NOT RELEVANT: Meanwhile, Kuwait remained on a state of alert in the wake of violence that left 14 people dead as security forces continued the hunt for a number of fugitives after killing eight of them and capturing 14 others.

21. RELEVANT: A Kuwaiti army spokesman said last Thursday that two Kuwaiti soldiers were to stand trial on suspicion of plotting attacks on US and other foreign forces in the country, which served as the main launchpad for the war.

22. NOT RELEVANT: Around 25,000 US soldiers are stationed in Kuwait, which is used as a transit point for US and other coalition troops headed for Iraq.

23. NOT RELEVANT: Staunch US ally Kuwait last week raised its state of alert almost to the maximum, boosting security around the country in the biggest show of force since the March 2003 launch of the US-led war in Iraq.

24. RELEVANT: Kuwaiti security forces have detained up to eight soldiers suspected of planning to attack US forces in the emirate, the Arab Times reported Tuesday.

25. NOT RELEVANT: There are some 25,000 US troops stationed in staunch Washington ally Kuwait, which serves as a transit point for coalition forces moving in and out of Iraq.

26. NOT RELEVANT: Security for US military convoys, using the emirate as a passage to neighbouring Iraq, has been boosted with more Kuwaiti police cars accompanying the convoys.

27. NOT RELEVANT: Kuwait 's Prime Minister Sheikh Sabah al-Ahmad al-Sabah warned Monday that Islamist violence which has rocked the emirate over the past month could spread to other oil-rich Gulf Arab states.

28. NOT RELEVANT: The embassy called on all US citizens to exercise caution, maintain a low profile, and avoid areas where Westerners are known to congregate.

29. RELEVANT: Investigation in the case of soldiers suspected of having intentions to attack the coalition forces has been completed, Mulla said in a statement reported by the state-run KUNA news agency.

30. NOT RELEVANT: We have begun redeploying our forces north and east of Gaza City to prevent any violations or attacks against Israel, a senior official told AFP on condition of anonymity.

31. NOT RELEVANT: The US embassy warned its citizens Monday that assailants in a black car driving around Kuwait planned to randomly attack Westerners, in a message posted on the embassy website.

32. NOT RELEVANT: It reminded American citizens of the potential for further terrorist actions against US citizens abroad, including in the Gulf region.

33. NOT RELEVANT: Earlier on Monday, Kuwait began to ease tight security measures introduced almost two weeks ago when the emirate raised its state of alert almost to the maximum in the biggest show of force since the March 2003 launch of the US-led war in Iraq.

34. RELEVANT: Security for US military convoys in Kuwait has been boosted, with more Kuwaiti police cars accompanying the convoys.

35. NOT RELEVANT: Americans who encounter suspicious vehicles matching this description should quickly but safely move away and contact the Kuwaiti police emergency number at 777, it added.

36. RELEVANT: Al-Qaeda -linked militants who over the past month fought four bloody gunbattles with Kuwaiti security forces had plotted to kidnap and execute US soldiers and Westerners, a newspaper reported Saturday.

37. NOT RELEVANT: Security for US convoys using the emirate as a rear base has been boosted with reinforced Kuwaiti police escorts and tighter controls on other traffic on the road.

38. NOT RELEVANT: Oteibi was arrested on the spot while militants in the car drove away and were being hunted by police, the ministry said.

39. NOT RELEVANT: Last month, the US embassy in Kuwait warned that it had credible information that terrorist groups were preparing to carry out attacks in the emirate in the near future and urged its citizens to be vigilant.

40. NOT RELEVANT: The US embassy warned December 15 it had credible information that terrorist groups were preparing to carry out attacks in the emirate in the near future.

GNG025 LIST FACTS ABOUT EVENT [The shut down of the Cernavoda nuclear power plant]
location constraint: Romania
source date constraint: 24 August – 01 December 2003

1. RELEVANT: By Friday, that had raised water levels to 2.5 meters 8.25 feet the minimum level needed for the cooling system to function, said Ionel Bucur, the general manager at the plant.

2. NOT RELEVANT: The Medzamor plant, 30 kilometers 20 miles west of the capital Yerevan, has one working Soviet-made reactor that supplies 30-40 percent of Armenia 's electricity.

3. RELEVANT: A nuclear power plant was shut down Sunday because a record drought left insufficient water to cool down the reactor.

4. NOT RELEVANT: The plant 's two 1,000-megawatt units are undergoing final trials required before they start commercial power generation.

5. RELEVANT: If it rains in Western and Central Europe, the increased water takes 25 days to reach Cernavoda, he told national radio, adding that the hot summer had led to a seven-percent increase in energy demands.

6. NOT RELEVANT: The Danube 's low level also has left taps in Cernavoda, a town of 20,000, coughing up brackish water unfit for drinking since Tuesday.

7. RELEVANT: With the plant supplying more than 10 percent of the country 's energy needs, the move fed fears of price hikes.

8. RELEVANT: Its the first time the plant in Cernavoda, some 200 kilometers 125 miles east of Bucharest, has encountered water level problems since it opened seven years ago.

9. NOT RELEVANT: The task force said it was looking into whether the voltage collapse may be tied to a lack of reactive power a portion of the energy moving through power lines that is essential for keeping proper voltage levels and balance.

10. RELEVANT: A nuclear power plant resumed operations on Friday, weeks after it was shut down because Romania 's severe drought had depleted water levels in the nearby Danube River that were needed for cooling.

11. NOT RELEVANT: Under the agreement, a subsidiary of Russia 's Unified Energy Systems, or UES, will oversee all payments made between Armenian electricity consumers and the power plant and will be responsible for paying Russian fuel suppliers.

12. NOT RELEVANT: He said UES subsidiary Inter UES will be responsible for the safe and uninterrupted operation of the nuclear power plant, which supplies 30-40 percent of Armenia 's electricity.

13. RELEVANT: The unprecedented dropping of water levels in the Danube in recent days made us decide to close the reactor at Cernavoda, said Dan Ioan Popescu, the minister for economy and commerce.

14. RELEVANT: A record drought forced the closure of a Romanian nuclear plant Sunday because of lack of water to cool the reactor.

15. RELEVANT: Popescu told the AM Pres news agency that the government had set up a crisis group and was trying to locate alternate sources of power.

16. RELEVANT: The Danube River at Cernavoda village, where the reactor is located, fell to a depth of less than three meters 10 feet on Saturday, down from its usual level of almost seven meters 23 feet.

17. NOT RELEVANT: The Danube 's low level also has left taps in Cernavoda, a town of 20,000, coughing up brackish water unfit for drinking since Tuesday.

18. RELEVANT: The heat wave gripping Europe, described by experts as one of the worst in 150 years, is slowly moving southeastward after killing what is believed to be thousands of people across the continent and shrinking major rivers to record-low levels.

19. NOT RELEVANT: The U.S. Energy Department has announced a contract for two American companies to oversee construction of the two coal-burning power plants.

20. RELEVANT: The plant was turned off on Aug. 24, after water levels dropped as low as 1.5 meters nearly 5 feet down from its usual level of almost seven meters 23 feet.

21. RELEVANT: The water level in the Danube River at Cernavoda village, where the reactor is located, fell to a depth of less than three meters 10 feet on Saturday, down from its usual level of almost seven meters 23 feet.

22. RELEVANT: The heat wave gripping Europe, described by experts as one of the worst in 150 years, is slowly moving southeast after killing what is believed to be thousands of people across the continent and shrinking major rivers to record-low levels.

23. NOT RELEVANT: Critics in Austria argue the plant should be shut down, while Czech authorities insist the plant is safe.

24. RELEVANT: The Canadian -designed plant has four reactors, but only one is in use.

25. RELEVANT: The Danube had its lowest flow of any August in Romania since measurements began in 1840, and a nuclear power plant was shut down a week ago because the drought left insufficient water to cool down the reactor.

26. RELEVANT: It is clear that energy produced from coal and fuel oil is much more expensive than other means, and we dont exclude energy hikes, he said.

27. RELEVANT: The plant supplies more than 10 percent of Romania 's electricity and closure prompted fears of a price hike.

GNG041 LIST FACTS ABOUT EVENT [a ferry crash] (location constraint: Canada)

location constraint: Canada

source date constraint: 01 February 2005 – 30 June 2005

source constraint: broadcast speech documents

1. NOT RELEVANT: October 19: Seven die after a wooden ferry sinks in a waterway in the northeastern Sunamganj district.

2. NOT RELEVANT: The capsizing of a ferry at the weekend, with 116 people confirmed dead by Monday and scores believed missing, is the latest in a string of ferry disasters that have killed thousands of people in Bangladesh.

3. NOT RELEVANT: On the same day a second ferry carrying a bridal party sinks in northern Kishoreganj district killing 52.

GNG044 LIST FACTS ABOUT EVENT [hunger strikes by Palestinians in Israeli jails]

source date constraint: 01 August 2004 – 31 August 2004

1. NOT RELEVANT: The Prison Authority, the Israeli today one of the more restrictions on hunger strike which started his Palestinian prisoners earlier today, according to the Radio Israel reported.

2. RELEVANT: It seemed thousands of detainees in the prisons of the Israeli Sunday on hunger strike for an improvement in their conditions of detention in prisons whiff Wayshil in south Israel Wahdarim north of Tel Aviv.

3. NOT RELEVANT: The Director of the Information Department of the Arab League Mahmoud Abdel Aziz, in a press statement following the meeting that the meeting held at the request of Palestine to mobilize public opinion against the Israeli practices in the prisons and the movement of international solidarity in this issue with a view to put pressure on Israel to abide by the rules of international law adequate in this regard and formulating a plan Arab action on the international scene regarding this issue.

4. NOT RELEVANT: Began at least 70 Palestinians representing different Palestinian national and Islamic factions and political forces today Saturday will go on a hunger strike in solidarity with the prisoners of Palestinians in prison.

5. RELEVANT: The Jordanian news agency that the president of the Federation of Women the Jordanian safe Alza bi revealed during their meeting, Millah at the Headquarters of the Red Cross to the demands of the Palestinian prisoners and the Jordanian fair and must be the response and the international pressure to respond to the international conventions and Human Rights expressed its concern about the Israel in its policy towards the Wamilha that there would be an international response to their demands focused on the issue of the prisoners, without charge or trial.

6. NOT RELEVANT: The Israeli Laflar spokesman of the Prisons Department told Agence France Presse that was involved in the strike in civilian prison, affirming that the strike did not extend to military detention.

7. NOT RELEVANT: He said that the strike and collective hunger is aimed at spoiling the Israeli policy aimed at preventing prisoners Security of the planning of terrorist attacks in their cells, saying that the strike organized by the Islamic Resistance Movement, HAMAS and Islamic Jihad.

8. RELEVANT: More than in the person of today in the West Bank to express their support for the detainees Palestinians who carry on a hunger strike in Israeli jails.

9. RELEVANT: The bodies in a statement made by the during the sit-in to the President of the Assembly of the International Red Cross in Jordan, Millah all international and humanitarian organizations and the International Committee of the Red Cross of pressure on Israel to respect and implement the fourth Geneva Convention and international humanitarian law with regard to the prisoners of war in Israeli prisons and the pressure for the release of prisoners are all in their release Alasirat and children under the age.

10. RELEVANT: The demonstrators, the slogans calling for the release of detainees and of releasing the prisoners of the brave.

11. RELEVANT: The statement pointed to the conditions of the families in the prisons of the Israeli that made the lives of the prisoners of death Watwaziyah and paid by violations of them to be coexistence with him and kept silent about the and the prisoners to fight the bowel, for the sake of dignity and decent life.

12. NOT RELEVANT: It was suspended nearly 700 detained in prison Jilbwa in northern Israel their strike today after the received promises to improve the conditions of the prison administration.

13. RELEVANT: For his part, said the Majdlawi one of the leaders of the Popular Front for the Liberation of Palestine that the activities of the hunger strike will escalate during the coming few days, denounced the statements made by minister of internal security threat Tsahi hanegby, in which he renewed its position not to accept the demands of the prisoners of Palestinians in prison.

14. NOT RELEVANT: The Palestinian officials today Saturday that about eight thousand Palestinians will begin a hunger strike in order to improve the conditions of their detention while the Israeli authorities in advance of any compromise.

15. NOT RELEVANT: The; prison al-khayam announced the end of last week their approval for the suspension of the strike which they mid- August after approved the prison administration to enter into negotiations with them on their demands and allowing them to contact with their strike in the other prisons.

16. NOT RELEVANT: The protesters demanded in a statement handed him to a representative of the Red Cross to intervene for humanitarian organizations to the Israeli authorities in order to improve the conditions of detention of Palestinians in prison.

17. NOT RELEVANT: He said Alnadi in a statement that a number of prisoners in prisons and al-khayam Ramla Watlmund Wajlbu and the Negev desert of prisoners and detainees in isolation Alanfaradi in Sajn Il- rmalh and other children in Telmond joined b 1700 prisoners in prisons Shatta and Nafha Walsb Wahdarim who entered their hunger strike on his fourth.

18. RELEVANT: Source said Arab official here today that the Council of the League of Arab States held here today held an emergency meeting at the level of permanent delegates to discuss the grave situation and the deterioration of the prisoners and detainees the Palestinians and the Arabs in the prisons.

19. NOT RELEVANT: The channel Al-Jazeera satellite channel that the suspension of the strike came to the prison authorities to part of the demands made by the prisoners at the beginning of their strike 0

20. NOT RELEVANT: He added that the emergency meeting to consider the conditions of more than prisoners and detainees Palestinian Arab in Israeli prisons were exposed to torture and

humiliation and psychological Wajidi and racist policies, has resulted in the death of some events and permanent disabilities among those heroes children and prisoners.

21. RELEVANT: The statement pointed out that 3500 other detention centres of the Israeli army Badawa today a solidarity with the prisoners strikers in the central prisons include the strike An food today for one day and the start tomorrow in the province of prisons and the return to the hunger strike Friday one day.

22. NOT RELEVANT: The 800 Palestinian prisoners in prison today, Friday, and even the two their hunger strike, which began mid- August, half of the prisoners the Palestinians nearly eight thousand detained in prisons, also announced the club of the prisoner.

23. RELEVANT: He pointed out the prisoners in contact with the prisoners and the supporters of the prisoner they feel that the management of the prison beguiled and procrastinating in the negotiation of the response to their demands.

24. RELEVANT: And in solidarity with the prisoners on hunger strike, started at least 70 Palestinians representing different Palestinian national and Islamic factions and political forces Saturday strike open for food, said Arab member of the Knesset 48/218 Israeli parliament that the Palestinians inside Israel would announce general strike on Wednesday next day of solidarity with the prisoners of the Palestinians.

25. NOT RELEVANT: It seemed about 500 Palestinian prisoners out of eight thousand detainees in the prisons of the Israeli hunger strike on Sunday to improve prison conditions, as well as to the prison authorities.

26. NOT RELEVANT: The Radio Israel quoted an as saying following a meeting of Ladirab prisoners that Israel would not accept any of the demands of the prisoners.

27. RELEVANT: And the flow of demonstrators took to the streets of Hebron, the meat and Tubas in the West Bank and Gaza city, carrying banners calling for the release of detainees and of releasing the prisoners of the brave.

28. RELEVANT: And half of 8000 detained in Israeli prisons in the movement of hunger strike, which started the one to improve the conditions of their detention.

29. NOT RELEVANT: The Israeli Laflar spokesman of the Israeli to reporters that the - meter arrest took part in the strike, saying that the strike did not extend to military detention.

30. RELEVANT: Systems thousands of Palestinians massive demonstrations today in the cities of the West Bank and Gaza Strip, to express their support for the detainees Palestinians on hunger strike in Israeli jails.

31. NOT RELEVANT: He underlined that the prison authority is preparing for the possibility of an extension of the strike for several months.

32. NOT RELEVANT: The about 1500 security has Badwa hunger strike this morning, Sunday, in prison Dishil Wanafiha Wahdarim to improve living conditions.

33. NOT RELEVANT: It seemed hundreds of about eight thousand Palestinian detainees in the prisons of the Israeli one on hunger strike in an attempt to improve the conditions of their detention.

34. NOT RELEVANT: The Al-Aqsa Martyrs battalions from the Fatah movement, one of its fighters in the Palestinian territories to execute more Al alamliat quality and martyrdom and focus on the kidnapping of Israeli soldiers and civilians in solidarity with the detainees Palestinians who started the day on hunger strike in prison.

Appendix C

Answers to GALE Year 1 Go/No-Go Evaluation Template 8 Questions

The answer prsented in this appendix were created using the automatically created *prosecution* domain template (Section C.1) and shallow semantic network (Section C.2).

C.1 Answers Created Using the Automatically Induced Domain Template

GNG015 DESCRIBE THE PROSECUTION OF [Saddam Hussein] FOR [crimes against humanity].
source date constraint: none
source constraint: none

1. NOT RELEVANT: He told reporters questions Saddam Hussein is accused of crimes committed in Iraq and they are normal crimes is.

2. NOT RELEVANT: He said that the persons involved in the suffering of Iraq are going to face justice in front of the Iraqi court specializing in crimes against humanity, genocide and war crimes.

3. NOT RELEVANT: Alawi said in a press conference, the government has officially asked the prisoners of the Iraqi tried for the crimes committed against the Iraqi people, and will be handed over Saddam Hussein and other symbols of the former regime to eliminate the Iraqi tomorrow.

4. RELEVANT: The same source said that Saddam Hussein is accused of crimes against humanity in seven cases.

5. NOT RELEVANT: The statement was attached with a copy of the letter signed by the wife of Saddam Hussein, and his daughters Rana and Raghad and solution to yesterday, the removal of the defence of President Saddam Hussein and responsibility for any action or action could affect the President Saddam Hussein.

6. NOT RELEVANT: She criticized the newspaper judge of the Court strongly that it is a weak point in the court Flashkhasith Walathiqaftah legal Walakhabrth qualify him for questioning people Kasdam Hussein Ajntah policy and refined experiments and made that nature of its bloody in the worst conditions Makabra in forcing the positions host it would have been better for the court to choose the age and deeper and more courageous and better-educated and sold the precise verbatim in his career judge who selected him, and in the central police of Iraqi many of them with the specifications.

7. NOT RELEVANT: The views of newspapers in accordance with the trends represented by some as Saddam Hussein as a weak Murtbka and trying to seek the sympathy of the court, while others were of the view that Saddam Hussein was allowed Waliqadi is accused.

8. RELEVANT: He was asked the spokesman about the statements made by Saddam Hussein to his appearance for the first time today before the Court where it described Bush as base, and tried to its representative for the purpose of elections Bush, that Saddam Hussein will continue to say many things.

9. NOT RELEVANT: The refusal of the Iraqi President Saddam Husseins signing of the indictment against him, which includes seven charges, as well as advocated the invasion of Kuwait in August, according to a senior official in the Special Criminal Court told Agence France Presse.

10. NOT RELEVANT: The newspaper, The Independent, she felt that the examination of Saddam in the technical aspects of and had not been successful but the returns of the wanted by miniaturization Saddam Wollalah and demonstrate the accused salary, which considers rope

around the corner of his neck, contending that the trial service more damaged by Wazhrth strong while I want the court to show its weak given Kasira.

11. NOT RELEVANT: In a challenge to the judge, the President of the former Iraqi signing the indictment, in the absence of his lawyer, terming the court to the scene of the election campaign of the American President George Bush.

12. NOT RELEVANT: The body of the shortcomings of yesterday that addressed the Iraqi court through the Bar Association, and unilaterally and without approval or even knowledge of the defense team asked to be the only counsel to Iraqi President Saddam Hussein without naming or placement of any lawyers of the defence.

GNG016 DESCRIBE THE PROSECUTION OF [Ali Hassan al-Majid] FOR [genocide].
source date constraint: none
source constraint: none

1. NOT RELEVANT: Iraqi Information Minister Mohammed Saeed al-Sahhaf denied al-Majid was killed, the Arabic-language station Al-Jazeera said.

2. NOT RELEVANT: Ali Hassan al-Majid, one of the most brutal members of President Saddam Husseins inner circle, was apparently killed by an airstrike on his house in Basra, British officials said Monday.

3. NOT RELEVANT: Human rights groups had called for al-Majids arrest on war crimes charges when he toured Arab capitals last January seeking to rally support against mounting U.S. pressure on Saddams regime.

4. NOT RELEVANT: During April 1991 peace talks in Baghdad, the Kurdish delegation leader, Jalal Talabani, told al-Majid that more than 200,000 Kurds lost their lives in the Iraqi campaign.

GNG017 DESCRIBE THE PROSECUTION OF [Khaled Jubran] FOR [terrorism].
source date constraint: none
source constraint: none

1. NOT RELEVANT: He was in a very bad mood the whole time, and he told me and other prisoners that he was beaten by intelligence officers during interrogation, testified Jubran, who is still in custody.

GNG033 DESCRIBE THE PROSECUTION OF [Zeljko Maksimovic] FOR [the murder of Bosko Buha].
source date constraint: none
source constraint: none
No answer sentences are selected.
GNG034 DESCRIBE THE PROSECUTION OF [Djordje Sevic] FOR [war crimes].
source date constraint: none
activity date constraint: 15 September 2003 – 01 December 2003
source constraint: none

1. RELEVANT: The four members of a notorious Serb paramilitary group known as Avengers who fought in Bosnia's 1992-95 war had each been sentenced to up to 20 years in prison for the abductions near the village of Sjeverin, close to Serbias border with Bosnia.

C.2 Answers Created Using a Shallow Semantic Network

GNG015 DESCRIBE THE PROSECUTION OF [Saddam Hussein] FOR [crimes against humanity].
source date constraint: none
source constraint: none

1. NOT RELEVANT: The appearance of Saddam Hussein 's court Iraqi special accused of committing crimes against humanity, reactions contradictory in Iraq.

2. NOT RELEVANT: He said that the persons involved in the suffering of Iraq are going to face justice in front of the Iraqi court specializing in crimes against humanity, genocide and war crimes.

3. NOT RELEVANT: He also said that Saddam Hussein and his aides would enjoy full rights, which have not been provided by the former regime to his victims, adding that the trial will be public and fair until conviction, adding that they have the right to have a lawyer to defend them and legal advice.

4. NOT RELEVANT: And on the legitimacy of the court said, We jurists Lanumah ; Man to the courts and the court set up by Paul Bremer the ruling of the U.S. to Iraq in June last year and

we do not recognize the legitimacy of the special courts Wakmhamin prefer to try Saddam in light of the Iraqi government elected and hostility and prefer to be held before the end of the Iraqi normal not extraordinary.

5. NOT RELEVANT: The newspaper union the mouthpiece of the Patriotic Union of Kurdistan PUK headed by Jalal Taliban that the trial of Saddam Wa wanah and channel 7 crimes against humanity is concerned, 3 of which are related to the crimes against our people the Kurds and satisfaction with the people of Kurdistan.

6. NOT RELEVANT: The newspaper, The Independent, she felt that the examination of Saddam in the technical aspects of and had not been successful but the returns of the wanted by miniaturization Saddam Wollalah and demonstrate the accused salary, which considers rope around the corner of his neck, contending that the trial service more damaged by Wazhrth strong while I want the court to show its weak given Kasira.

7. NOT RELEVANT: On the other hand, the Kuwaiti Information Minister Mohammad Abul Hassan, head of Saddam Hussein as a war criminal committed many crimes against Kuwait and its people and to Iraq and its people, he should be tried for all crimes Saddam without exception, of which the crime of invasion.

8. NOT RELEVANT: The authority warned of a lawyer did not obtain agency of President Saddam Hussein, she said, pointing out that such agency must come from the defence and with the written approval of the family of President Saddam Hussein.

9. NOT RELEVANT: And Saddam in the first of July before the investigating judge 264/93 charges of committing crimes against humanity.

10. NOT RELEVANT: The investigating judge Iraqi today for committing crimes against humanity to the Iraqi president Saddam Hussein in seven cases, said a senior official at the Iraqi Special Criminal Court told Agence France Presse.

11. NOT RELEVANT: He went on to say, I am looking forward to the day when will face elements of the former regime to the officials of the justice and absolute for the crimes committed against the Iraqi people.

12. NOT RELEVANT: It went on to say that the joy of Saddam 's trial Wa wanah not because they will suffer the fate of the unfortunate for their crimes but have its influence in the future because it is a milestone in the future of the region where its history, to a great extent to the

crimes against humanity of killing, destruction and devastation, the trial was important in the history of Iraq in the region and the world.

13. NOT RELEVANT: He told reporters questions Saddam Hussein is accused of crimes committed in Iraq and they are normal crimes is.

14. NOT RELEVANT: The Iraqi president was arrested in last December 13 in a village near his birthplace in Tikrit, such as for the first time before the judge of the last Thursday and 7 charges, including the use of weapons Alkmiawih against the Kurds and the invasion of Kuwait and crimes against humanity.

15. NOT RELEVANT: The Iraqi newspapers published here today Saturday interest in the trial of Iraqi President Saddam Hussein and described it as a trial.

16. NOT RELEVANT: Said Kamal Hamdoun, the chairman of the Iraqi lawyers that the syndicate received today Wednesday a request from the defence of Iraqi President Saddam Hussein, asking for the approval of the agency for defending Saddam.

17. RELEVANT: He was asked the spokesman about the statements made by Saddam Hussein to his appearance for the first time today before the Court where it described Bush as base, and tried to its representative for the purpose of elections Bush, that Saddam Hussein will continue to say many things.

18. NOT RELEVANT: The president of the former Iraqi last December and for the first time today to the special Iraqi court that he was accused of committing crimes against humanity in seven major issues.

19. NOT RELEVANT: In the pictures of the president of the former Iraqi and 11 of his senior aides the front pages have been allocated some newspapers several pages to publish the details of the trial the charges leveled against Saddam Hussein and his top aides, in addition to the reactions of the trial and the Arab and international levels.

20. NOT RELEVANT: The body of the shortcomings of yesterday that addressed the Iraqi court through the Bar Association, and unilaterally and without approval or even knowledge of the defense team asked to be the only counsel to Iraqi President Saddam Hussein without naming or placement of any lawyers of the defence.

GNG016 DESCRIBE THE PROSECUTION OF [Ali Hassan al-Majid] FOR [genocide].

source date constraint: none

source constraint: none

1. NOT RELEVANT: If his death is confirmed, Majid would be the most senior member of the Iraqi government known to be killed since the launch of the US-led war March 20 aimed at toppling Saddam 's regime.

2. NOT RELEVANT: People in the area of the village continue to die of cancer and suffer from asthma, sterility and miscarriages, according to Kurdish doctors.

3. NOT RELEVANT: Majid was charged by Saddam with the final solution to the Kurdish problem and orchestrated the genocide against the Kurdish people from February to August 1988, he said.

4. NOT RELEVANT: He dropped scheduled stops in Jordan and Egypt, both U.S. allies.

5. NOT RELEVANT: During April 1991 peace talks in Baghdad, the Kurdish delegation leader, Jalal Talabani, told al-Majid that more than 200,000 Kurds lost their lives in the Iraqi campaign.

6. NOT RELEVANT: Jackson said a body that was thought to be his was found along with that of his bodyguard and the head of Iraqi intelligence services in Basra.

7. NOT RELEVANT: When it became clear the United States would launch a war to topple Saddam, Majid was appointed governor of southern Iraq to organize the defense of the region – and to ensure that the mass uprising urged by the coalition did not materialize.

8. NOT RELEVANT: Both brothers were lured back to Iraq in February 1996 and killed on their uncle 's orders, together with several other family members.

9. RELEVANT: Majid could be among the first of Saddam 's 11 jailed henchmen to go on trial for war crimes, crimes against humanity and genocide, any of which could result in a sentence of death by hanging or firing squad.

10. NOT RELEVANT: Associated Press writer Maamoun Youssef contributed to this report from Cairo, Egypt.

11. NOT RELEVANT: Saddam Hussein 's notorious senior aide Ali Hassan al-Majid, known as Chemical Ali for ordering the gas attack that killed about 5,000 Kurdish villagers in 1988, was reported dead Monday after a coaltion air strike on his villa.

12. NOT RELEVANT: The guerrillas retreated to the surrounding hills, leaving behind women and children, and on March 16, 1988, jets swooped over the town and for five hours sprayed mustard gas and nerve agents, including Sarin.

13. NOT RELEVANT: New York-based Human Rights Watch, in a report earlier this year, called for the arrest and prosecution of Majid, saying he was responsible for the deaths or disappearances of around 100,000 non-combatant Kurds when he put down their revolt.

14. NOT RELEVANT: Ali Hassan al-Majid, one of the most brutal members of President Saddam Hussein 's inner circle, was apparently killed by an airstrike on his house in Basra, British officials said Monday.

15. NOT RELEVANT: Central Command said Saturday that two coalition aircraft had attacked Majid 's residence with laser-guided munitions. One of Majid 's bodyguards was confirmed dead shortly afterward, an official at the base here said Sunday.

16. NOT RELEVANT: Before the start of the war, he was named governor of Iraq 's southern province with the goal of defending the region and ensuring that the mass uprising called for by the United States and Britain would not materialize.

17. NOT RELEVANT: Al-Majid apparently was killed Saturday when two coalition aircraft used laser-guided munitions to attack his house in Basra.

18. NOT RELEVANT: Majid had experience in Basra, having crushed a Shiite Muslim-led uprising in southern Iraq that erupted, with US encouragement but little assistance, in the wake of the 1991 Gulf War.

19. NOT RELEVANT: Believed to be in his fifties, al-Majid led a 1988 campaign against rebellious Kurds in northern Iraq in which whole villages were wiped out.

20. NOT RELEVANT: In 1988, as the 1980-88 Iran- Iraq war was winding down, he commanded a scorched-earth campaign to wipe out a Kurdish rebellion in northern Iraq.

21. NOT RELEVANT: Egypt refused to receive him and the Jordanian government denied a visit was ever planned.

22. NOT RELEVANT: In the video, which was shown on several Arab TV networks, al-Majid was seen executing captured rebels with pistol shots to the head and kicking others in the face as they sat on the ground.

23. NOT RELEVANT: Jackson said the apparent discovery of al-Majid 's body was one of the reasons the British decided to move infantry into Basra, because they hoped that resistance in the southern Iraqi city might crumble with the top leadership gone.

24. NOT RELEVANT: Human rights groups had called for al-Majid 's arrest on war crimes charges when he toured Arab capitals last January seeking to rally support against mounting U.S. pressure on Saddam 's regime.

25. NOT RELEVANT: Amid fears of a Kurdish alliance with Tehran, Iraqi jets bombed the agricultural town of Halabja near the Iranian border to flush out fighters from the Patriotic Union of Kurdistan.

26. NOT RELEVANT: His nephew and Saddam 's son-in-law, Lt. Gen. Hussein Kamel, was in charge of Iraq 's clandestine weapons programs before defecting in 1995 to Jordan with his brother, Saddam Kamel, who was married to Saddam 's other daughter.

27. NOT RELEVANT: Al-Majid replied that the figure was exaggerated and the dead were not more than 100,000, according to Arab press reports.

28. NOT RELEVANT: Al-Majid also has been linked to the bloody crackdown on Shiites in southern Iraq after their uprising following the 1991 Gulf War.

29. NOT RELEVANT: Dozens of Kurds lay lifeless in front of their homes many with blood pouring out of their noses as they tried in vain to flee the attacks.

30. NOT RELEVANT: But at the US Central Command 's forward planning base here, a senior US commander stressed there was no hard evidence yet Majid was dead.

31. NOT RELEVANT: He has been involved in some of Iraq 's worst crimes – including genocide and crimes against humanity, said Kenneth Roth, Human Rights Watch 's executive director.

32. NOT RELEVANT: It is believed to have been the biggest gas attack carried out against civilians.

33. NOT RELEVANT: Prior to that, he served as governor of Kuwait during Iraq 's seven-month occupation of its neighbor in 1990-1991, an invasion that led to the Gulf War.

34. NOT RELEVANT: After Iraq 's 1991 Shiite Muslim uprising was crushed, Iraqi opposition groups released a video they said had been smuggled out of southern Iraq.

35. NOT RELEVANT: He had been dubbed Chemical Ali by opponents for ordering a 1988 poison gas attack that killed thousands of Kurds.

36. NOT RELEVANT: British Defence Secretary Geoff Hoon said there were strong indications that Majid, who is President Saddam 's cousin, was killed in a coalition raid on the southern city of Basra, which was overrun by thousands of British troops overnight.

37. NOT RELEVANT: Chemical Ali was also accused of brutalities when he served as Iraq 's military governor during the seven-month occupation of Kuwait that was ended by the US-led 1991 conflict.

38. NOT RELEVANT: While it would have been more satisfying to see al-Majid answer for his crimes in an international war crimes tribunal, the hundreds of thousands of victims of his genocide campaign must be finding some solace in his death, Barham Salih, prime minister of the Kurdistan regional government in Sulaymaniyah said in a statement.

39. NOT RELEVANT: Some 5,000 people were killed, three-quarters of them women and children.

40. NOT RELEVANT: British Maj. Andrew Jackson of the 3rd Battalion Parachute Regiment told The Associated Press that his superiors had reported the death of the man who was Saddam 's first cousin, entrusted with defending southern Iraq against invading coalition forces.

41. NOT RELEVANT: He has been made by Saddam as a military governor of Kuwait, which is in the province is 19 after the occupation of Iraq Ljarth South in August 1990 that Saddam replaced him three months for fear that the reputation of described by the West to the brutal position allies of the Kuwaitis who were their exile in Saudi Arabia.

42. NOT RELEVANT: Later, he boasted about the attacks, including the March 16, 1988, poison gas strike on the village of Halabja, where an estimated 5,000 people died.

43. NOT RELEVANT: Syria and Lebanon ignored international calls to arrest al-Majid when he visited in January.

GNG017 DESCRIBE THE PROSECUTION OF [Khaled Jubran] FOR [terrorism].

source date constraint: none

source constraint: none

1. RELEVANT: The presiding judge, Col. Fawn Buqour, sentenced them to five years in jail with hard labor but promptly reduced the sentences to 2 1/2 years owing to the circumstances of the case and to give the accused an opportunity to repent.

2. NOT RELEVANT: Condemned France Friday strongly the deadly attack, saying, the spokesman of the Ministry of Foreign Affairs of the French head of Ladsu that France strongly condemns such evil killed a large number of victims, including Ayatollah Mohammad Baqir al-Hakim, personal Iraqi Shiite leader.

3. NOT RELEVANT: Chechen rebels have attacked Russian military convoys in the restive region, killing 14 troops, a senior Russian official was quoted as saying Saturday.

4. NOT RELEVANT: About 50 Israeli armoured vehicles raided a Palestinian refugee camp Saturday, one day after Israeli Prime Minister Ariel Sharon accepted the Road Map peace plan.

5. NOT RELEVANT: Iraq 's US administrators Saturday began paying wages to state employees for the first time since the fall of Baghdad, as a contested election in the oil-rich northern city of Kirkuk put in place a city council.

6. NOT RELEVANT: A body invited National Health in cooperation with the Council students of the Faculty of Health Sciences at the University of America to Discuss The other face of the aggression on Iraq, studies on the health effects, environmental and psychological, Robert Hanna, Malik chattila Elie Karam, run by Dr Iman Nuwayhad, at 5.00 pm 8, in a hall in the University of America

7. NOT RELEVANT: The Bilathyu that the site was targeting is a place special importance to the Shiites, and once again show the need for the international community to show solidarity and the fight against terrorism and sought to establish security and stability in Iraq.

8. NOT RELEVANT: Authorities allege he sent threatening e-mails to officials from Jordan and other countries because they prevented mujahedeen, or holy fighters, from attacking Americans and Israelis.

9. NOT RELEVANT: Iran sees no need to immediately revive a dialogue with the United States, following the latest round of talks to discuss who should govern postwar Iraq, Iran 's top diplomat was quoted on Saturday as saying.

10. NOT RELEVANT: Kuwaiti Interior Minister Sheikh Mohammed Khaled al-Sabah affirmed Kuwait has taken the necessary precautions and measures in anticipation of any terror act and pledged readiness to strike the heads of terror cells.

GNG033 DESCRIBE THE PROSECUTION OF [Zeljko Maksimovic] FOR [the murder of Bosko Buha].

source date constraint: none

source constraint: none

1. RELEVANT: A special prosecutor on Tuesday charged seven men with killing police Gen. Bosko Buha near a popular Belgrade Danube River restaurant on June 10, 2002.

2. NOT RELEVANT: A criminal group accused of carrying out Djindjic 's murder will go on trial in November.

3. NOT RELEVANT: Another criminal group accused of carrying out Djindjic 's murder in downtown Belgrade will go on trial in November before the special court.

4. NOT RELEVANT: Following Djindjic 's assassination, Serbia 's leadership launched a major crackdown on organized crime that flourished under former President Slobodan Milosevic.

5. RELEVANT: One of the defendants, Dragan Ilic, claimed in court that he was tortured in custody following his arrest last year.

6. NOT RELEVANT: Following the assassination of Serbian Prime Minister Zoran Djindjic in March, authorities in the Balkan republic launched a major crackdown on organized crime that flourished when Slobodan Milosevic was president.

7. RELEVANT: A special prosecutor 's office has been given broad powers to investigate scores of unsolved murder cases, including Djindjic 's assassination.

8. RELEVANT: An underworld group charged with assassinating a ranking police officer went on trial Tuesday before a special court established to prosecute organized crime.

9. RELEVANT: The defendants, including the man who allegedly pulled the trigger, all denied the charges at the start of the trial, which is open to the public.

10. RELEVANT: According to the indictment, the seven defendants also had plans to murder several top Serbian politicians, including Djindjic.

11. NOT RELEVANT: Veteran Li Nan and Yao Ming, NBA 's top pick draft last year, contributed 16 points of 25 points together.

12. NOT RELEVANT: Two army colonels and a lieutenant colonel were arrested after they allegedly were paid for revealing secrets to an intermediary working for Russia 's secret service, the Beta news agency said.

13. RELEVANT: Serbia 's new court tasked with prosecuting organized crime opened its first trial Tuesday, charging an underworld group with assassinating a senior police officer.

14. RELEVANT: According to the indictment, the defendants also planned to murder several top Serbian politicians, including Djindjic.

15. RELEVANT: The special prosecutor charged the alleged criminal group with killing police Gen. Bosko Buha near a popular Belgrade Danube River restaurant on June 10, 2002.

16. NOT RELEVANT: Lefteris Alexiou / Alex Jakupovic, Greece, bt Dejan Petrovic/ Nenad Zimonjic, Serbia and Montenegro, 6-7 6-8, 6-3, 6-2, 6-4

17. RELEVANT: Serbia 's first trial targeting organized crime is seen as a test of the republic 's judicial system.

18. RELEVANT: The trial was conducted by a recently established special prosecutor 's office, which has been given broad powers to investigate scores of unresolved murder cases in the Balkan republic, including the assassination of Serbian Prime Minister Zoran Djindjic in March.

GNG034 DESCRIBE THE PROSECUTION OF [Djordje Sevic] FOR [war crimes].
source date constraint: none
activity date constraint: 15 September 2003 – 01 December 2003
source constraint: none

1. NOT RELEVANT: While families of the Sjeverin victims welcomed the retrial Monday, Kandic said she feared a dangerous trend by the Supreme Court after it threw out another landmark war crimes sentence last week that of 20 years imprisonment for a Serb police officer found guilty of executing 14 ethnic Albanian civilians during the 1999 Kosovo war.

2. NOT RELEVANT: Lukic, 37, is suspected to have been involved in that case also.

3. RELEVANT: The four belonged to a dreaded paramilitary Serb group known as Avengers, which fought Muslims during the 1992-1995 Bosnian war.

4. NOT RELEVANT: Kosovo is a Serbian province but it has been under UN and NATO control since June 1999 after NATO air strikes forced Serbian forces to withdraw and end a crackdown on the separatist ethnic Albanian majority.

5. NOT RELEVANT: Milosevic 's government is believed to have sponsored dozens of paramilitary groups that were active during the Bosnian and other wars in former Yugoslavia in the 1990s.

6. RELEVANT: The sentencing at the Belgrade District Court caps years of effort by relatives of the victims and nongovernment organizations to bring the high-profile case to court.

7. RELEVANT: Four former Serb paramilitary soldiers were convicted Monday on war crimes charges and sentenced to up to 20 years in prison for the killing of 16 Muslim civilians in Bosnia in 1992.

8. NOT RELEVANT: Djindjic made many enemies by orchestrating Milosevic 's extradition to the U.N. war crimes tribunal in The Hague, Netherlands, and by declaring war on organized crime.

9. NOT RELEVANT: Lukic, allegedly hiding somewhere in Serbia, is also wanted by The Hague tribunal in connection with crimes committed in eastern Bosnia during the war.

10. RELEVANT: Later in the day, they were transported to the eastern Bosnian town of Visegrad, where they were shot or stabbed to death and thrown into the Drina River.

11. RELEVANT: On the day of their abduction, they were traveling to work in Bosnia.

12. RELEVANT: The two who appeared in court, Djordje Sevic and Dragutin Dragicevic, received 15 and 20 years in jail respectively.

13. RELEVANT: The two, who went into hiding before the trial started last January, were sentenced to 20 years in prison on charges of war crimes against civilians.

14. RELEVANT: Serbia 's democratic authorities, who ousted Milosevic in 2000, are struggling to escape the legacy of ethnic conflict and his ruinous regime.

15. NOT RELEVANT: Cvjetan, who pleaded not guilty at the trial, was the first Serb policeman to have been tried for war crimes committed during the 1998-99 war in Kosovo, which left about 10,000 people dead, mostly Albanians.

16. RELEVANT: According to the verdict, the first for war crimes since the parliament appointed a special war crimes prosecutor last July, the paramilitaries stopped a bus traveling from Serbia to neighboring, wartorn Bosnia in October, 1992, forcing 15 Muslim men and a woman to get off.

17. RELEVANT: The four members of a notorious Serb paramilitary group known as Avengers who fought in Bosnia 's 1992-95 war had each been sentenced to up to 20 years in prison for the abductions near the village of Sjeverin, close to Serbia 's border with Bosnia.

18. RELEVANT: According to the verdict, the first for war crimes since the parliament appointed a special war crimes prosecutor last July, the paramilitaries in October 1992 stopped a bus traveling from Serbia to neighboring war-torn Bosnia, forcing 15 Muslim men and a woman to get off.

19. NOT RELEVANT: Milosevic is now on trial at the U.N. war crimes court in The Hague, Netherlands for genocide and other atrocities committed in the Balkan wars of the 1990s.

20. RELEVANT: Although the massacre took place before the eyes of many witnesses, former Serbian authorities under ex- President Slobodan Milosevic refused to deal with the case.