The Hong Kong Polytechnic University

Department of Computing

Investigations on Temporal-Oriented Event-Based Extractive Summarization

Wu Mingli

A Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy

June, 2007

CERTIFICATE OF ORIGINALITY

I hereby declare that this thesis is my own work and that, to the best of my knowledge and belief, it reproduces no material previously published or written, nor material that has been accepted for the award of any other degree or diploma, except where due acknowledgement has been made in the text.

_____(Signed)

Wu Mingli (Name of student)

Abstract

Automatic summarization aims to produce a concise summary of source documents by identifying the focused topics in documents. Normally, topics are represented by some essential events. Topics may evolve or shift over time. Tracking the trend of the topics requires anchoring events on the time line. Unfortunately, both events and their associated time features are not well studied in previous work. Investigating event-based and temporal-oriented summarization techniques are primary objectives of this study. As a matter of fact, the salience of contents could hardly be evaluated from single point of view. Exploiting a framework which can effectively integrate multiple impact factors is another objective.

We define events by "action" words as well as associated named entities. Events weave documents into a map built either on event instances or event concepts. Relevance between events is exploited to identify important events. To utilize temporal information associated to events, it is necessary to extract and normalize temporal expressions. We investigate rule-based approaches for these tasks. Two statistical measures are employed to evaluate the significance of events based on their temporal distributions. Sentence selection is a complicated process. Therefore we explore various features including surface, content, event and relevance features under a learning-based classification framework. Event-based and temporal-oriented approaches are incorporated as features into this framework.

The contributions of this study are listed as follows. (1) Event-based summarization approaches are proposed. They achieve competitive results when compared with successful word-based approaches. (2) Temporal concepts are introduced into event-based summarization and temporal information is found crucial to summarization on documents which contain evolving topics. (3) An adaptive leaning-based framework is developed to incorporate various types of features. (4) A system for temporal expression extraction and normalization is implemented. It is an effective tool not only practical for document summarization, but also for many other applications.

Publications Arising From the Thesis

[1] **Wu**, **M.**, Li, W., Lu, Q. and Wong, K.F. 2007. Exploiting surface, content and discourse features for extractive summarization. In Proceedings of IEEE International Conference on Natural Language Processing and Knowledge Engineering (NLPKE-2007), pages 234-241, Beijing, China.

[2] **Wu**, **M.**, Li, W., Lu, Q. and Wong, K.F. 2007. Event-based summarization using time features. In Proceedings of the Eighth International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2007), pages 563-574, Mexico City, Mexico.

[3] Liu, M., Li, W., **Wu, M.** and Lu, Q. 2007. Extractive summarization based on event term clustering. In Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL-2007), Prague, Czech Republic.

[4] Liu, M., Li, W., **Wu, M.** and Hu, H. 2007. Event-based extractive summarization using event semantic relevance from external linguistic resource. In Proceedings of the 6th International Conference on Advanced Language Processing and Web Information Technology (ALPIT-2007), Luoyang, China.

[5] Li, W., **Wu**, **M.**, Lu, Q. and Wong, K.F. 2006. Integrating temporal distribution information into event-based summarization. International Journal of Computer Processing of Oriental Languages, vol. 19, no. 2-3, pages 201-222.

[6] **Wu, M.** 2006. Investigations on event-based summarization. In Proceedings of the Student Research Workshop of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, pages 37-42, Sydney, Australia.

[7] **Wu**, **M.**, Li, W., Lu, Q. and Wong K.F. 2006. Chinese temporal expression extraction and normalization. Submitted to ACM Transactions on Asian Language Information Processing.

[8] Li, W., Xu, W., **Wu**, **M.**, Yuan, C. and Lu, Q. 2006. Extractive summarization using inter- and intra- event relevance. In Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (COLING/ACL-2006), pages 369-376, Sydney, Australia.

[9] Xu, W., Li, W., **Wu**, **M.**, Li, W. and Yuan, C. 2006. Deriving event relevance from the ontology constructed with formal concept analysis. In Proceedings of the Seventh International Conference on Intelligent Text Processing and Computational Linguistics (CICLING-2006), pages 480-489, Mexico City, Mexico.

[10] **Wu**, **M.**, Li, W., Chen, Q and Lu, Q. 2005. Normalizing Chinese temporal expressions with multi-label classification. In Proceedings of 2005 IEEE International Conference on Natural language Processing and Knowledge Engineering (NLPKE-2005), pages 318 -323, Wu Han, China.

[11] **Wu**, **M.**, Li, W., Lu, Q. and Li, B. 2005. CTEMP: A Chinese temporal parser for extracting and normalizing temporal information. In Proceedings of the Second International Joint Conference on Natural Language Processing (IJCNLP-2005), pages 694-706, Jeju Island, South Korea.

[12] Li, B., Li, W., Lu, Q. and **Wu, M.** 2005. Profile-based event tracking. In Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR-2005), pages 631-632, Salvador, Brazil.

Acknowledgments

This thesis could not have been done without the help and cooperation of many people, and it is my great pleasure to take this opportunity to thank them.

First and foremost, I would like to express my deepest thanks to my chief supervisor Dr. Wenjie Li and co-supervisor Prof. Qin Lu, for being a consistent source of support and encouragement. I could not imagine having another better advisor or mentor for me. Without their knowledge and perceptiveness, I would never have finished my Ph.D. study. I gratefully acknowledge them for their instruction, help and encouragement in my study and life.

It would be my great pleasure to thank Dr. Robert Luk and Dr. James Liu, for their great efforts, valuable comments and excellent advices to improve the quality of my thesis and my previous research report. Other great excellent staffs whom I would like to express my deep gratitude are Dr. Ruifeng Xu and Dr. Baoli Li, my close friends, for the continuous support and kind help.

At last, I would like express my deepest appreciation to my father Shizhang Wu and my mother Yongzhen Chen, for their endless love and unwavering support. This thesis is dedicated to them.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Concepts

Text summarization is the process of distilling the most important information from a source (or sources) to produce an abridged version for a particular user (or users) and task (or tasks) [Mani and Maybury 1999]. **Text summaries** are reduced texts which convey important information, products of text summarization. In this study, summarization and summaries are restricted in text summarization and text summaries, respectively. Examples of original text and the corresponding summary are shown as bellows:

Example of original text: "A small airplane crashed into a government building in heart of Milan, setting the top floors on fire, Italian police reported. There were no immediate reports on casualties as rescue workers attempted to clear the area in the city's financial district. Few details of the crash were available, but news reports about it immediately set off fears that it might be a terrorist act akin to the Sep. 11 attacks in the United States. Those fears sent U.S. stocks tumbling to session lows in late morning trading."

Example of summary: "A small airplane crashed, setting the top floors on fire, Italian police reported. News reports set off fears that it might be a terrorist act akin to the Sep. 11 attacks."

According to different classification standards, summaries can be divided into different type pairs, such as {**generic** summary / **query-based** summary} and {**extract** / **abstract**}. Generic summaries are aimed at a particular broad readership community. Traditionally **generic** summaries written by authors or professional abstractors serve as surrogates for full-

text. On the other hand, **query-based** summaries are tailored to a question or the requirements of particular users. Extract and abstract are two different kinds of summaries. An **extract** is a summary consisting entirely of material copied from the input. It means that an extract at a condensation rate will take certain material in the document. In contrast, an **abstract** is a summary at least some of whose material is not present in the input. The summarization process to generate extract and abstract is named **extraction** and **abstraction**, respectively. Generic extraction is investigated in this study.

1.2 Background and Motivations

With the growing of electronic text information nowadays, it is inefficient for users to browse a great number of individual documents. To deal with large amount of information, people expect to see the important part of them only. Automatic text summarization involves condensing a document or a document set to produce a fair summary based on the power of computer techniques. The object is to help users catch the important contents of the original document(s) with bearable time costs. Summarization is widely useful, as it can be employed alone or combined with other applications, such as information retrieval. As machine generated summaries can not match human written ones at present, it is necessary and meaningful to investigate automatic summarization in the background of information explosion.

According to the number of documents to be summarized, automatic text summarization approaches are classified as multi-document summarization and single-document summarization. As multiple related documents are about a same or similar theme, overlapped information may be contained. If similar contents are identified as important, multi-documents summarization systems should remove the redundant part. However, single-document summarization systems can omit this consideration. Multi-document summarization approaches have been widely used and receive more attention recently. We focus on them in this study.

The previous research on text summarization can date back to [Luhn 1958] and [Edmundson 1969]. In the following periods, two alternative summarization approaches have been suggested in the literature, named abstraction and extraction. Abstraction typically needs comprehensively "understand" and then paraphrases the salient concepts across documents. Limited by current natural language processing techniques, it is confined in specific domains. Extractive summarization, on the other hand, selects sentences which contain the most significant concepts in documents. Significance can be evaluated statistically or empirically. Although the performance of extractive summarization can not satisfy the requirement of human, it is rather effective and applicable.

Extractive summarization extracts part of document(s) based on some weighting scheme, in which different features are exploited, such as position in document, term frequency, and key phrases. Recently extraction approaches may also employ machine learning approaches to decide which sentences or phrases should be extracted. According to features and labels of sentences in training data, sentences in testing data can be decided whether they should be extracted or not based on the input feature values. Some researchers devote their efforts to investigate it and achieve preliminary success in different applications.

Previous extractive approaches identify the important content mainly based on terms. Bag of words may be not a good representation for meanings, as there is so little useful information associated with these words. A predefined template is a better choice for representation of documents. However it is domain-dependent and need much effort to create and fill it. This tension motivates us to seek a balance between effective implementation and deep understanding. Certain kind of auto-built semi-structured representation may convey more meaningful contents and be easier to achieved, such as event.

Recently researchers [Filatova and Hatzivassiloglou 2004] define events by "action" words as well as associated named entities including person names, organization names, locations and times. Given the sentence "Yasser Arafat on Tuesday accused the United States of threatening to kill PLO officials", "accused", "threatening" and "kill" are identified

as an event terms, while “Yasser Arafat”, “United States”, “PLO” and “Tuesday” are event elements. Event represented in this way contains more precise and structured information than bag-of-words, and meanwhile could be constructed with less effort than predefined template. Based on this definition and event frequency, Filatova and Hatzivassiloglou [2004] show promising results of event-based summarization.

Events are not stand-alone. They can weave contents of documents into a map built either on event instances or event concepts. An instance is one occurrence of an event in original documents, while a concept is the set of instances which are same lexically. When people mention a same event at different positions of documents, they refer to the same concept. When they mention different instances/concepts at a same sentence, these instances/concepts are related under a certain context. In previous researches, the relevance between event instances or concepts isn’t exploited well. It motivates this research to focus on whether the relationship between event instances or concepts can improve the summarization procedure. Intra- and inter- event relevance can be easily introduced in the concept-based approach, while the instance-based approach differentiates instances of a event concept so that further investigation can be conducted later, such as utilizing distribution of event instance on the time line.

Document(s) may contain descriptions on a topic at different periods. For example, the theme of a cluster of documents is “fires in California” and the model summary mentions the fires in 1926, 1977, 1985, 1987 and 1990. This observation originates the investigation on whether taking event distributions on the time line into account can improve the quality of summaries of these kinds of clusters in the context of event-based summarization. It is assumed that if an event occurs frequently in a specific period and rarely in other periods, then it is important in this period. By extracting important events in different periods, we can collect a synopsis of documents. To testify the assumption, events should be anchored on the time line according to their corresponding normalized temporal attributes, and the weights of

events should be evaluated according to their temporal distribution. Sentences in documents are then selected according to the weights and temporal distribution of the events.

To utilize temporal information associated to events, it is necessary to extract and normalize temporal expressions. Temporal information is defined as the knowledge about time point or duration. This information is crucial for both temporal reasoning and anchoring events on the time line. Temporal expressions are defined as chunks of text which convey direct or indirect temporal information, which includes dates, times of day, positions of dates, durations, set-denoting expressions, event-anchored expressions, and so on. To retrieve the useful temporal information contained in these temporal expressions, we need to identify the extents of temporal expressions in raw text and then represent temporal information according to certain standard. The two tasks are called temporal extraction and temporal normalization, respectively. In literature, rule-based approaches are reported rather applicable and successful.

The key point of extractive summarization is not sentence or paragraph selection, but how to select. It is inadequate to evaluate sentences according to a single criterion. Typically, the effects of different types of features on sentence significance are integrated by linear combinations. The drawback of such schemes is that the weight parameters of the features have to be tuned experimentally. It is time-consuming and may overlook the best combination of parameters. One solution to this problem is involving learning mechanism to make the combination optimal, at least in a trainable way.

The performance of learning based approaches depends heavily on input features. Recently different sentence features are explored individually, but most of them are not incorporated into a learning based approach and their functions are not well studied when they are combined together. At the same time, event relevance in a sentence and relevance between sentences are not exploited in previous learning based approaches. Therefore we conduct this study on learning-based framework and various features. Event-based and/or temporal-oriented approaches are incorporated as features into this framework.

1.3 Objectives

In view of the limitations and problems of previous summarization approaches, the main objective of this research is to investigate approaches to improve the performance of text summarization. The generated summaries should be used reliably to help users catch key points of documents. The focus of this research is on finding effective representation for documents contents and effective weighting schemes for the representation in the context of extractive summarization. In addition, to incorporate evidences from multiple points of view, a unified summarization framework should be designed and evaluated.

To achieve this objective, investigations are divided into four parts.

1. **Investigate semi-structured representations of document contents and design effective weighting schemes based on these representations.**

   Representation of document contents is the basis for text summarization. All weighting schemes depend on certain representation, such as bag-of-words. Motivated by the tension between deep representation and bearable cost, we plan to investigate a balanced means -- event. Event is defined as event terms and associated event elements, such as person names, organization names, times and locations. Event instance is the basic unit to convey useful information from texts. At the same time, set of lexically same instances – event concept is a represent unit at high level. The strength and shortness of the two representations will be studied in this research.

   In real text, event instances or concepts are not independent. We plan to consider both intra-event and inter-event relevance for summarization. Intra-event relevance measures how an event term itself is linked with the associated event elements. This is a kind of direct relevance as the connections between actions and arguments are established from the text surface directly. On the other hand, we plan to consider how an event term (or a named entity involved in an event) is associated to another event term (or another named entity involved in the same or different

events). After connect relevant instances or concepts, an event map is built. A map weighting algorithm will be employed to evaluate the importance of each node in this map. Finally, weight scores of sentences will be calculated based on weights of events contained in them.

2. **Extract and normalize temporal expressions for temporal-oriented summarization.**

   Temporal information is useful for many natural language processing applications, such as text summarization, information extraction and machine translation. To reflect the temporal distribution of events, we need to achieve values of temporal attributes of events. Temporal expressions are chunks of text which convey direct or indirect temporal information, which is crucial for anchoring events on the time line. Therefore temporal expressions have to be extracted from real text and then normalized according to certain guidelines, such as TimeX2 [Ferro et al. 2004; Gerber et al. 2004]. Temporal expressions are classified into dates, times of day, positions of dates, durations, set-denoting expressions, event-anchored expressions, etc. Temporal attributes include VAL, MOD, SET, ANCHOR_VAL and ANCHOR_DIR. As temporal expressions are confined in a limited domain, rule-based approaches will be investigated. These approaches are time-consuming, but normally they can achieve high performance. Rule-based approaches will be employed to identify different kinds of temporal expressions and give the values for different temporal attributes.

3. **Design a temporal-oriented event-based summarization approach to reflect the event trends in documents.**

   Some clusters of documents consist of descriptions about topics at different periods, and then the human generated summaries present the event trends of these topics. By utilizing this characteristic, the summarization performance is expected to be improved. Events will be anchored on the time line according to their temporal

attributes, and then effective weighting schemes should be investigated. It is assumed that if an event occurs frequently on certain periods and rarely on other periods, this event is important at the specific periods. Then different statistics of events will be calculated to reflect the strength that an event is associated with different periods, i.e. the importance of that event at different periods. The weights of sentences depend on weights of events contained in them.

4. **Design a unified summarization framework to incorporate evidences from multiple points of view.**

It is difficult evaluate the importance of a sentence from a single criteria. A unified summarization framework will be built to incorporate different evidences which are from different importance evaluation aspects, such as position of sentence, number of frequent words, sum of the weight of events in a sentence, and so on. Sentence extraction can be transfer to a classification problem, whether a sentence should be extract or not. Typical classification algorithms can be employed to give the decision based on the input sentence features. The length of final summary is fixed and the length of extracted sentences judged by the classifier may not exactly mach this limitation. A re-ranking algorithm should be designed to order sentences so that they can be picked up one by one according to this order.

Since sentence features can influence the final summary heavily, we plan to investigate basic surface features, content features a sentence may represent and the features indicating the relevance among sentences. While surface and content features are about extrinsic and intrinsic aspects of a sentence itself, relevance features describe the strength of sentence relatedness. Event features are also investigated to reflect how important events contained in sentences are. They can be derived from our event-based summarization approaches. Under this framework, all evidence can be unified and a rather reliable decision on sentence extraction will be outputted.

1.4 Contributions

The contributions of this study are listed as follows. (1) We propose event-based summarization approaches. They not only achieve competitive results when compared with successful word-based approaches, but also provide potential way to sentence compression. This is an important step toward abstraction. (2) We introduce temporal concepts into event-based summarization and suggest that temporal information is crucial to summarization on documents which contain evolving topics. (3) We develop an adaptive leaning-based framework to incorporate various types of features. The framework could achieve optimal weight parameter combination, compared with those based on experimentally tuned parameters. (4) We implement two systems for English and Chinese temporal expression extraction and normalization. They are effective tools not only for document summarization, but also for many other applications, such as question answering and machine translation.

1.5 Organization

The thesis is organized as follows. Chapter 2 gives details of background information and related work. Chapter 3 presents instance-based and concept-based event summarization approaches. The relevance between events is exploited by building an event map and then evaluating the importance of each node. Chapter 4 describes the investigations on temporal expression extraction and normalization. Details about kinds of temporal expressions and attributes are given. Chapter 5 proposes event summarization based on event distribution on the time line. Summarization approaches with and without temporal information exploiting are compared. Chapter 6 describes a unified summarization framework which can incorporate various sentence features, such as importance from event summarization and number of frequent words. Then the framework output the decision based on all these features and pick up sentences according to their re-ranked order. Finally Chapter 7 gives the conclusions of this study.

Chapter 2

Literature Review

This research focuses on improving text summarization by investigating different document representation and weighing schemes. To reflect and exploit shifting topics on the time line, we also explore temporal information processing techniques, mainly temporal expression extraction and normalization. Therefore, the related works can be divided into two parts: summarization and temporal information processing. They are described as follows in detail.

2.1 Summarization

Summarization is the procedure to distill important information from document(s) for users. It has been widely investigated in the past [Mani and Maybury 1999]. Mani [2001] and Sparck-Jones [1999] give a good introduction. The previous research on text summarization can date back to [Luhn 1958] and [Edmundson 1969]. In the following periods, some researchers focus on extractive summarization, as it is effective and simple. Others investigate abstractive summarization, but their work is highly domain-dependent or preliminary investigation. Query-based summarization involves query representation and anchoring query relevant contents in documents. It is highly related to information retrieval, another research subject.

The key problems are how to represent contents of texts and how to identity the important part. Recently, semi-structured representations between bag-of-words and predefined template are explored, such as event. To catch topics evolving over time, temporal information processing is incorporated into summarization. It is difficult to

evaluate the importance of sentence/paragraph by single criteria. Therefore machine learning techniques are employed to combine multiple features. The literature on summarization is discussed from these aspects.

Multi-document summarization, which is contrast to single-document summarization, may consider redundancy in different documents. Various similarity measures are used. A common approach is to measure similarity between pairs of sentences and then use clustering to identify themes of common information [Radev, Jing, and Budzikowska 2000a; Marcu and Gerber 2001]. Similar sentences will not be included in summaries. Another possible way is to measure the similarity of a candidate passage to that of already-selected passages and retain it only if it contains enough new (dissimilar) information [Radev, Jing, and Budzikowska 2000a; Carbonell and Goldstein 1998].

2.1.1 Abstractive Summarization

Radev et al. [2002] give a brief introduction about abstraction methods. At this early stage of the research on summarization, we categorize any approach that does not use extraction as an abstractive approach. Abstractive approaches have used information extraction, compression, ontological information and information fusion.

Information extraction approaches can be characterized as "top-down", since they look for a set of predefined information types to be included in the summary. For each topic, users predefine templates of expected information types, together with recognition criteria. For example, an earthquake template may contain slots about location, earthquake magnitude, number of casualties, etc. The summarization engine must then locate the desired pieces of information, fill them in slots, and generate a summary with the results [Dejong 1978; Rau and Jacobs 1991]. This method can produce high-quality and rather accurate summaries, although it is restricted in certain domains only.

Compressive summarization results from approaching the problem from the point of view of language generation. Using the smallest units form the original document, Witbrock

and Mittal [1999] extract a set of words from the input document and then order the words into sentences using a bigram language model. Jing and McKeown [1999] think human summaries are often constructed from the source document(s) by a process of cutting and pasting document fragments, which are then combined and regenerated as sentences in summaries. Hence a summarizer can be developed to extract sentences, reduce them by dropping unimportant fragments, and then use information fusion and generation to combine the remaining fragments.

Other researchers focus on the reduction process. In an attempt to learn rules for reduction, Knigh and Marcu [2000] train a system to compress the syntactic parse trees of sentences in order to produce a shorter but still maximally grammatical version. Ultimately, this approach can likely be used for shortening two sentences into one.

True abstraction involves taking the process one step further. Abstraction involves recognizing that a set of extracted passages together with something new, something that is not explicitly mentioned in the source, and then replacing them in the summary with the new concept(s). The requirement that the new material not be in the text explicitly means that the system must access to external information of some kind, such as an ontology or a knowledge base, and be able to perform inference [Hahn and Reimer 1999]. Since no large-scale resource of this kind yet exists, abstractive summarization has not progressed beyond the proof-of-concept stage.

Lehnert [1982] propose summarization strategy that builds upon the prior research [DeJong 1979]. She set up three primary affect states: positive event, negative event, and mental event, and four links between the events: motivation, actualization, termination and equivalence. Given the events and links, she designs 15 legal pair-wise configurations (problem, success, failure, resolution and so on) which are building blocks for more complex stories. Some plot units are pivotal in driving inferences about other plot units, so identification of pivotal units is very important in summarization. She regards an event as a dot and connects events with causal links, and then she can get a plot-unit graph structure.

After identify the pivotal unit the algorithm will give a base-line summary from it. At last, augment the base-line with information from plot units related to the pivot one. She does not implement the algorithm and leaves some questions open, but it may provide a potential foundation to generate narrative summaries.

Hahn and Reimer [1999] propose using a formal terminological knowledge model as the presentation of contents of documents. Text summarization is considered as an operator-based transformation process and knowledge representation structures are mapped into conceptually more abstract structures forming a text summary. The authors illustrate their system implemented in their model on information technology reviews and legal reports. They describe a three-step procedure for summarization. First, a list of salience operators is applied to paragraphs. Second, topic descriptions are determined and aggregated over paragraphs as appropriate. Third, generalization operators are applied across the topic descriptions to create a hierarchical text graph.

Mckeown et al. [1995] describe an approach to summary generation that folds information from multiple facts into a single sentence. They present two systems using this approach. STREAK system is used to generate summaries about basketball games. The system creates a draft of facts and then applies revision rules. PLANDOC system is about telephone network planning activity, which employs discourse planning. It looks ahead to group facts together by conjunction and deleted repetitions.

Traditional language generation systems include a content planner and a surface sentence generator. Content planner selects information in proposition size chunks. Surface sentence generator decides whether the sentence is an interrogative or declarative expression. Then it selects the main verb of the sentence and maps elements in the fact to verb arguments. At last it enforces syntactic constraints and produces the words in linear order. Being different with the traditional approaches, authors' method distinguishes essential and optional content, using revision strategies to pack information into a sentence.

Maybury [1995] gives attention to the summarization of structured information sources, such as data base or event collection. The author attempts to identify the key event among many simple event messages. He employs these factors: event frequencies, frequencies of relations between events and domain specific importance. After determining which events would remain in the summary, the system condenses the events using aggregation. If two events have the same agent, time or location, they will be combined. Finally, all events are grouped by the mission they related to. Following aggregation, the overall organization of the resulting narrative is planned using a text generator [Maybury 1991; Maybury 1992] that organize the report first by topic, and then temporal relation under a topic.

2.1.2 Extractive Summarization

2.1.2.1 Overview

Radev et. al [2002] give a good introduction about extraction approaches. Despite the beginning of research on alternatives to extraction, most work today still relies on extraction of sentences from original documents to form a summary. The majority of early extraction approaches focus on the development of relatively simple surface-level techniques that tend to signal important sentences/paragraphs in the source text. Although most systems use sentences as units, some work with paragraphs. A set of features is computed for each sentence/paragraph, and ultimately these features are normalized and summed. Sentences/paragraphs with the highest scores are extracted and returned as the extract.

Early extraction approaches for sentence extraction compute a weight score for each sentence based on features such as position in the text [Baxendale 1958; Edmundson 1969], word and phrase frequency [Luhn 1958] and key phrases (e.g., “it is important to note”) [Edmundson 1969]. Recent approaches use more sophisticated techniques to decide which sentence should be extracted. These techniques often rely on machine learning approaches to identify important features, on natural language analysis to identify key sentences/paragraphs, or on relations between words rather than bags of words.

Approaches involving more sophisticated natural language analysis to identify important sentences/paragraphs rely on word relatedness or discourse structure. Some researchers use the degree of lexical connectedness between a paragraph and the remainder of the text; connectedness may be measured by the number of shared words, synonyms, or anaphora [Salton et al. 1997; Mani and Bloedorn 1997; Barzilay and Elhadad 1999a]. Others reward sentences/paragraphs that include topic words, i.e. words that have been determined to correlate well with the topics of interest to users or with the general theme of the source text [Buckley and Cardie 1997; Strzalkowski et al. 1999; Radev et al. 2000b].

Alternatively, a summarizer may reward paragraphs that occupy important positions in the discourse structure of the text [Marcu 1997b; Ono et al. 1994]. This method requires a system to compute discourse structure reliably, which is not possible in all genres. Teufel and Moens [2002] focus on it. They show how particular types of rhetorical relations in the genre of scientific journal articles can be reliably identified through the use of classification. Conroy and O'Leary [2001] has turned to the use of hidden Markov Models (HMMs) and pivoted QR decomposition to reflect the fact that the probability of inclusion of a sentence in an extract depends on whether the previous sentence has been included as well.

To identify redundancy in text documents, various similarity measures are used. A common approach is to measure similarity between all pairs of sentences and then use clustering to identify themes of common information [Radev et al. 2000a; Marcu and Gerber 2001]. Other approaches use information fusion techniques to identify repetitive phrases from the clusters and then combine the phrases into the summary [Barzilay et al. 1999b]. Mani et al. [1999] describe the use of human-generated compression and reformulation rules. Alternatively, systems measure the similarity of a candidate paragraph to already-selected ones and retain it only if it contains enough new (dissimilar) information. A typical approach of this kind is Maximal Marginal Relevance (MMR) [Carbonell et al. 1997; Carbonell and Goldstein 1998].

Ensuring coherence is difficult, because this in principle requires some understanding of the content of each paragraph and knowledge about the structure of discourse. In practice, most systems simply follow temporal order and text order. To avoid misleading the reader when juxtaposed paragraphs form different dates all say "yesterday", some systems add explicit time stamps [Lin and Hovy 2002]. Other system use a combination of temporal and coherence constraints to order sentences [Barzilay et al. 2001]. Recently, Jahna et al. [2002] have focused on discourse-based revisions of multi-document clusters as a means for improving summary coherence.

The cohesion relations in summarization include synonym/hyponym relations, repetition, adjacency in a phrase and co-reference, not just word-based vectors similarity. To extracting phrase, Mani and Bloedorn [1999] use a robust finite-state parsing based on patterns defined over part-of-speech tags. To extracting synonym and hyponym they used WordNet. In their text content map, a word is a node. There are adjacency links, same links (lexical or semantic), and phrase-building links. Then a document can be represented as a graph. Given a topic, they find topic-related text regions by weighting the graph and then select segments from the weighted graph. To summarize multiple documents, they use cross-document sentence extraction and cross-document sentence alignment. Their approach touches the surface of the problem.

2.1.2.2 Extraction without Exploiting Discourse Structure

Automatic summarization can date back to 1950's. In perhaps the first paper in this field, Lun [1958] describes a summarization system based on simple sentence extraction. The problem he faced is to give the importance of sentences and extract the most important sentences. Firstly, he establishes a set of "significant" words, whose frequencies are between a higher bound and a lower bound. It is assumed that high-frequency or low-frequency words are unimportant. Then he derives significance of a sentence based on the number of the occurrences of significant words in the sentence.

Edmundson [1969] also develops a primary summarization system. In addition to word frequency, he exploits cue phrases, words in titles and headings, sentence location and particular section. The principle of the word frequency is like the one first proposed by Luhn [1958], but the algorithm is not the same one. Cue phrases include positively relevant, negatively relevant and irrelevant ones. The cue dictionary is compiled on the basis of statistical data and refined by linguistic criteria. The final "cue" weight score for each sentence is the sum of the cue weight scores of its constituent phrases.

He supposes that an author conceives the title as circumscribing the subject of the document. Those sentences which contain words in titles, subtitles and headings should have more significance. The final "title" weight score for each sentence is the sum of the title weight scores of its constituent words.

Edmundson [1969] makes another hypotheses that sentences occurring under certain headings("introduction", "purpose", "conclusions") are positively relevant and topic sentences tend to occur very early or very late in a document and its paragraphs. In addition to assigning positive weight scores provided by the heading dictionary, the method also assigns positive weight scores according to sentence position in the text, i.e. in first or last paragraph, and as first or last sentence of a paragraph.

Finally, the relative weights from the four basic measures are parameterized in the linear function "$a_1C + a_2K + a_3T + a_4L$". The values of a_1, a_2, a_3 and a_4 can be specified as desired. By their observation, the method which exploited cue phrase, title words, and location seems to have the best performance.

Pollock and Zamora [1975] aim at automatically generating chemical abstract by sentence rejection algorithm, rather than sentence selection. Their sentence rejection is based on negative semantic words and some syntactic features, such as having no verbs in sentence. Word frequency here is used to modify the semantic code, for example, to make positive code less positive and make negative code less negative. As a final step, the summary is

automatically edited to delete certain non-substantive words and phrases which occur at beginning of sentences.

Brandow et al. [1995] describe a system that performs domain-independent condensation on commercial news. Their system contains four major constituents: statistical corpus, signature word selection, sentence weighting and sentence selection. When tf*idf for a given word is more than a threshold, they would regard this word as a signature word. In addition, they add the headline words into the signature words. Then each sentence's weight is computed by summing the weights of the individual signature words occurring in the sentence. Finally, they use some heuristics to extract some sentence from a document as the summary. This system is evaluated against a system that using only the first portion of the texts (leading-text) and the result is that the leading-text method is better.

After these prior studies on extractive summarization, researchers focus on effective features for extraction of sentence or phrase recently. Surface features are about extrinsic characteristics of a sentence, such as position of sentence, length of sentence, and whether the sentence is the first sentence of a document or paragraph. Content features are about intrinsic characteristics of a sentence, such as number of centroid words, signature words or frequent words in a sentence. Other features are from the point of view of event, such as number of persons, organizations, locations and times in a sentence. A combination function or machine learning algorithm will incorporate these features and give the importance of a sentence.

2.1.2.3 Extraction with Exploiting Discourse Structure

Boguraev and Kennedy [1997] try to represent a document by some phrases, which referring to the most prominent objects mentioned in the discourse. They obtain the phrases by a typical term identification algorithm. They use anaphora resolution method to establish crucial connections between text expressions that refer to the same entities. To produce summaries, they segment text using a variant of the approach proposed by Hearst [1994], which is a segmentation method relying on similarity between blocks of text based on

vocabulary overlap. They identify the “topic stamps” for a text segment as several highest ranked objects in it, according to the frequency of use and grammatical distribution of phrases in the text. To form a summary, the co-referential phrases associated with topic stamps are listed, along with some information from the surrounding context.

Barzilay and Elhadad [1997] also establish their summary system by sentence extraction. They separate the summarization procedure as four steps. The original text is segmented first. Then lexical chains are constructed and strong chains are identified. Finally significant sentences are extracted. They use Hearst’s algorithm on text segmentation [Hearst 1994] and build chains for each segment. Two chains are merged across a segment boundary only if they contain a common word with the same sense. The node candidate in a chain is noun or noun compound, and then they connect the nodes which have relations in the WordNet. Three kinds of relations are defined: extra-strong (between a word and its repetition), strong (between two words connected by a WordNet relation) and medium-strong when the distance of the link between the synsets of the words is longer than one. According to the idea of text cohesion, finally the algorithm will retain a best chain which has most links in it. Among all the lexical chains in a document, they select the strongest lexical chains as the topic indication by empirical methods. They find the following parameters are good predicators of the strength of a chain: the number of occurrences of the members of the chain, the number of distinct occurrences divided by the length. If the product of the two parameters satisfies their “criterion”, they regard the chain as a strong chain. To form summary according to the strong lexical chain, they select sentences by heuristics. They give three alternatives and each one will select one sentence for one strong chain.

In his paper, Marcu [1999] propose that the concepts of discourse structure can be used effectively in text summarization. Based on Rhetorical Structure Theory [Mann and Thompson 1988], Marcu [1997a; 1997b] describes a rhetorical parsing algorithm which receives unrestricted text and derives a rhetorical structure tree. In the rhetorical structure tree, each node represents a text unit, such as a clause or a sentence. After getting the

discourse structure tree, they determine a partial ordering on the elementary and parenthetical units of the discourse structure tree. A very simple way to induce such a ordering is computing a score for each text unit on the basis of the depth in the tree structure. At last, the summary algorithm selects some important text units according to their scores.

Strzalkowski et al. [1999] exploit an empirical observation that much of the written text displays certain regularities of organization and style, which they call Discourse Macro Structure. In order to produce a coherent summary they select paragraphs from the source document and assemble them into a mini-document within a DMS template. They think that certain types of text conform to relatively simple macro discourse structures, for example [Rino and Scott 1994] have shown that both physics papers and abstracts align closely with the "Introduction Methodology Results Discussion Conclusion" macro structure. Their selection criteria include word frequency, title words occurrences, noun phrase occurrences, paragraph location, proper name occurrences and cue phrase occurrences. Then a combination function will integrate scores which come from each selection measure. Finally they will trim the paragraphs and give the summary.

Teufel and Moens [1999] extend the KPC methodology [Kupiec et al. 1995] by classifying extracted sentences according to their rhetorical roles. They propose that the rhetorical roles included goal, achievement, background, method, etc. They regard abstract as an argumentative template, where its slots represent certain rhetorical roles. Then the system is trained to find suitable sentence to fill the slots. Their idea is more related to the structured abstracts [Hartley et al. 1996; Rennie and Glass 1991]. First, they extract sentences with a Bayesian classifier. Second, they classify the sentences according to one of the seven rhetorical roles by another Bayesian classifier. The features they employed include cue phrase, phrase about rhetorical roles, location, sentence length, title, word frequency and header (discussion, introduction, conclusion and so on).

Salton et al. [1997] characterize the text structure of a document by intra-document linkage pattern. They apply the knowledge of text structure to do automatic text

summarization by paragraph extraction. Every paragraph is represented by a text vector and they computed a similarity between every two paragraphs. If the similarity is smaller than a threshold, there is no "semantic" link between the two paragraphs. Then they extract the central paragraphs which have more links and present them to users. The performance of the system is acceptable but not perfect.

The approaches we have mentioned above are all extraction approaches. Although these methods generate functionally acceptable summaries, the performance is still very moderate. With the prevalence of machine learning methods, some researchers tend to improve the sentence extraction by machine learning based approaches.

2.1.3 Event-Based Summarization

Term-based extractive summarization [Luhn 1958; Edmundson 1969] represents the content of documents mainly by bag of words. Luhn [1958] establishes a set of “significant” words, whose frequency is between a higher bound and a lower bound. Edmundson [1969] collects common words, cue words and title/heading words from documents. Weight scores of sentences are computed based on type/frequency of terms. Sentences with higher scores will be included in summaries. Later researchers adopt tf*idf scores to discriminate words [Brandow et al. 1995; Radev et al. 2004]. Term-based approaches can not express exact meanings of documents and can just output applicable summaries. Therefore it needs to be improved further.

To represent deep meaning of documents, other researchers have investigated different structures. Barzilay and Elhadad [1997] segment the original text and construct lexical chains. They employ strong chains to represent important parts of documents. Marcu [1997a] describes a rhetorical parsing approach which takes unrestricted text as input and derives rhetorical structure trees. They express documents with structure trees. Dejong [1978] adopts predefined templates to express documents. For each topic, users predefine frames of

expected information types, together with recognition criteria. However, these approaches just achieve moderate results.

To balance the document representation between bag-of-words and deep structure (i.e., template [MUC-7 1998]), event [Filatova and Hatzivassiloglou 2004; Li et al. 2005; Liu et al. 2007a; Liu et al. 2007b] receives interests of researchers recently. Event-based summarization is first presented in [Daniel et al. 2003], who consider a news topic to be summarized as a series of sub-events according to human understanding of the topic. They determine the degree of sentence relevance to each sub-event by human judgment and evaluated six extractive approaches. They conclude that recognizing sub-events that comprise a single news event is essential for generating better summaries. However, it is difficult to break a news topic into sub-events automatically.

Later, atomic events are defined as the relationships between the important named entities [Filatova and Hatzivassiloglou 2004], such as participants, locations and times (which are called relations) through the verbs or action nouns labeling the events themselves (which are called connectors). They evaluate sentences based on co-occurrence statistics of named entity relations and the event connectors involved. They claim that the proposed approach achieves better results when compared with a conventional tf*idf term based approach. Apparently, named entities are key elements in their model. However, the constraints defining events seems quite stringent.

The application of dependency parsing, anaphora and co-reference resolution in recognizing events involves NLP and IE techniques more or less [Yoshioka and Haraguchi 2004; Vanderwende et al. 2004; Leskovec et al. 2004]. Rather than pre-specifying events, they extract (verb)-(dependent relation)-(noun) triples as events automatically and take the triples to form a graph merged by relations.

As a matter of fact, events in documents are related in some ways. Judging whether sentences are salient or not and organizing them in a coherent summary can take advantage from event relevance. Unfortunately, this is neglected in most previous work. Barzilay and

Lapata [2005] exploit the use of the distributional and referential information of entities to improve summary coherence. While they capture text relatedness with entity transition sequences, they conduct the summarization based on entities. We are particularly interested in relevance between events in this study and we conduct the summarization based on events and event relevance.

Extractive summarization requires ranking sentences with respect to their importance. Successfully used in Web-link analysis and more recently in text summarization, Google's PageRank [Page et al. 1998] is one of the most typical ranking algorithms. It is a kind of graph-based ranking algorithm to judge the importance of a node within a graph by taking into account the global information recursively computed from the entire graph, rather than relying on only the local node-specific information. The application of PageRank in sentence extraction was first reported in [Erkan and Radev 2004]. A graph can be constructed by adding a node for each sentence. Edges between nodes are established using inter-sentence similarity relations as a function of content overlap. The same idea is followed and investigated extensively in [Mihalcea 2005].

Yoshioka and Haraguchi [2004] go one step further toward event-based summarization. Two sentences are linked if they share same events, not same words. When evaluated on TSC-3, the approach favors longer summaries. In contrast, the importance of verbs and nouns constructing events is evaluated with PageRank as individual nodes connected by their dependence relations [Vanderwende et al. 2004; Leskovec et al. 2004].

2.1.4 Temporal-Oriented Summarization

Temporal information processing receives more attention than ever, such as at TERN 2004 [TERN 2004] and ACE 2005 [ACE 2005]. It can be employed to improve the existing approaches in text summarization, question answering and information extraction, etc. One way to improve an event-based summarization system is to discover the trends of events by exploiting event temporal distributions. Two fundamental issues in temporal information

processing are recognizing and normalizing temporal expressions in texts [Mani and Wilson 2000; Schilder and Habel 2001]. As one objective of this study is to investigate the application of temporal information processing in summarization, we simply normalize temporal expressions and assign the attribute values to the corresponding events manually first. In the future we plan to implement this procedure automatically.

The applications of temporal information in summarization have been investigated in the past, but most of them are based on publication dates. Given a sequence of news reports on certain topic, Allan et al. [2001] extract sentences with usefulness and novelty to monitor the changes. Usefulness is captured by considering whether a sentence can be generated by a language model created from the sentences seen to date. Novelty is captured by comparing a sentence to prior sentences. They report that it is difficult to combine the two factors successfully. Afantenos et al. [2005] discuss the techniques to summarize events happened synchronously, such as football matches reported at same times but from different sources. Relations between events (i.e., messages) are defined on the axis of time and information source. They are determined by comparing different messages with heuristic rules. However, they do not report the evaluation on summaries.

Other researchers exploit distribution of events on the time line by statistical measures. Swan and Allan [2000] aim at extracting and grouping important terms to generate "topics" defined by TDT 1998 [TDT 1998]. They employ statistics to measure the strength that a term is associated with a specific date. After filtering by a threshold, the significant terms are clustered into a few topics. Subjective evaluation shows the overlap between machine generated clusters and model topics is 86.7%. Lim et al. [2004] anchor documents on the time line by the publication dates, and then extract sentences from each document based on surface features. Time slots (dates) are used to extract high frequency words in each slot, and then to identify one topic sentence in the slot. Sentence weight is adjusted by these local high frequency words. Finally, global high frequency words form all topic sentences are used to adjust weights of sentences. They evaluate the system on Korean documents and report that

time can help to raise the percentage of model sentences contained in machine generated summaries.

Jatowt and Ishizuka [2004] investigate approaches to monitor the trends of dynamic web documents, which mean the different versions of same web documents. Based on distributions, terms are scored in order to identify whether they are popular and active. They employ a simple regression analysis on word frequency and time. A term's slop, intercept and variance are used to evaluate its importance. Then each sentence is weighted based on the sum of the weights of the terms that are contained in the sentence. Finally, the sentences with highest scores are extracted into a summary. Unfortunately, they don't report any quantitative evaluation results.

2.1.5 Learning-Based Summarization

The application of machine learning to summarization is pioneered by Kupiec et al. [1995], who develop a summarizer using a Bayesian classifier to combine features from a corpus of scientific articles and their abstracts. Myaeng and Jang [1999] adopt an algorithm like Kupiec et al. [1995], but they give the probability that a sentence should be in the summary, according to each independent feature, and then use Dempster-Shafer combination rule [Rich and Knight 1991] to calculate the belief that each sentence is contained in a summary.

Aone et al. [1999] design a summary system based on frequency approach, but they try to use some linguistic knowledge. They extract multi-word phrases and use them as the basic unit. Corpus knowledge is incorporated in three ways, by using a large corpus baseline database to calculate *idf* values for selecting signature words, by deriving collocations statistically from a large corpus, and by creating a word association index. They adopt a trainable Bayes classifier, which is like that of [Kupiec et al. 1995]. The features are sentence length, inclusion of signature words, sentence position in a document and sentence position in a paragraph.

Aone et al. [1999] and Lin [1999] investigate other forms of machine learning and its effectiveness. Machine learning has also been applied to individual features. For example, Lin and Hovy [1997] apply machine learning to the problem of determining how sentence position affects the selection of sentences. Witbrock and Mittal [1999] use statistical approaches to choose important words and phrases.

Machine learning approaches give decisions based on input features. These features influence summarization performance heavily. Teufel and Moens [1999] and Radev et al. [2004] report that position and length of sentences are effective surface features. They observe that sentences located at the beginning of a document most likely contain important information. They also find that important sentences are not likely to be too short. The role of content features is also widely acknowledged in the past. Luhn [1958] compiles a list of important words, whose frequencies are between higher and lower bound. Then the number of these words contained in a sentence is used to evaluate the importance of the sentence. Recently, other content indicating features are proposed, including centroid [Radev et al. 2004], signature terms [Lin and Hovy 2000] and high frequency words [Nenkova et al. 2006].

Radev et al. [2004] define centroid words as those whose average tf*idf are higher than a threshold. Meanwhile, Lin and Hovy [2000] identify signature terms that are strongly associated with documents based on statistics measures. Nenkova et al. [2006] later report that high frequency words are also crucial in finding focuses of documents. These content features are simply based on the statistics of uni-grams, except signature term. As bi-grams and tri-grams contain more information than uni-grams, it is reasonable to employ them as sentence features for summarization.

Document structure is another feature investigated in the context of summarization. Barzilay and Elhadad [1997] construct lexical chains and extract sentences based on the chains with higher scores. Marcu [1997a] applies a rhetorical parsing approach which takes unrestricted texts as input and derives the rhetorical structure trees. However, these structure-based approaches can only achieve moderate results. Dejong [1978], in contrast, predefine

templates to represent documents. This approach requires much effort to create templates and it can hardly be adapted in different domains.

Recently, event-based document representations receive much attention. While Filatovia and Hatzivassiloglou [2004] define event as actions (verbs/action nouns) and the associated named entities, Vanderwende et al. [2004] represent event by dependency triples. They then employ frequency of events or PageRank to evaluate sentence importance and achieve encouraging results. Li et al. and Wu [Li et al. 2006b; Wu 2006] later report that named entities are good indicators to locate the focuses of documents. It suggests that event features are worth to explore.

As a matter of fact, sentences in a document are connected in some way. Sentence relevance has been used as an alternative means to identify importance sentences. Erkan and Radev [2004], and Yoshioka and Haraguchi [2004] compare every pair of sentences. A web analysis approach, PageRank, is introduced to pick up the important sentences from the map built on sentence relevance. Promising results are reported in their papers.

2.2 Temporal Information Processing

Motivated by its potential applications, temporal information processing has attracted more attention recently than ever. A comprehensive review on recent trends has been given by Mani [Mani et al. 2004]. Research work in this area is classified into four classes: designing annotation schemes for temporal information representation [Ferro et al. 2004; Gerber et al. 2004; Sauri et al. 2004]; developing temporal ontology which covers temporal objects and their relationships [Allen 1984; Hobbs and Pan 2004]; identifying time-stamps of events or temporal relationships between events [Filatove and Hovy 2001; Li et al. 2001]; and extracting and normalizing temporal expressions in different languages [Ahn et al. 2005; Estela et al. 2002; Jang et al. 2004; Mani and Wilson 2000; Schilder and Habel 2001; Vazov 2001]. Temporal annotation, temporal ontology and temporal reasoning are not in the scope of this study.

Among the research work on temporal expression extraction and normalization, most are based on hand-written rules. Vazov [2001] extracts temporal expressions based on context constraints and regular expressions. Mani and Wilson [2000] normalize temporal expressions by hand-crafted rules. Schilder and Habel [2001] employ finite state transducers based on hand-written rules to extract and normalize temporal expressions. However, their evaluation is conducted on a small corpus. Jang et al. [2004] focus on Korean languages and the temporal expressions they cope with include fully-specified temporal expressions (e.g., April 3, 2000), context-dependent temporal expressions (e.g., "tomorrow" and "Thursday") and event-denoting temporal expressions (e.g. "after the election"). Building a rule based system is quite time-consuming.

Machine learning approaches are investigated for the task of temporal expression normalization. One potential application of learning approaches is the classification problem. In normalization, sometimes, we have to choose one interpretation of a temporal expression from several alternatives. Mani and Wilson [2000] develop a machine learning classifier to distinguish the specific and generic uses of "today". Ahn et al. [2005] evaluate different machine learning approaches (e.g. maximum entropy) for classification problems in normalization, such as whether a temporal expression refer to a time point or duration, whether a temporal expression refer to a point before or after the reference time. They all report promising results. Another application of learning approaches is normalizing temporal expressions directly. Jang et al. [2004] employ a rote learning approach to flag temporal expressions in testing documents, based on a dictionary which contains the temporal expressions from training documents. However, the results simply based on learning approaches have left much room to be improved.

Chapter 3

Event-Based Extractive Summarization

3.1 Chapter Overview

The amount of on-line documents grows fast with the expansion of the Internet. It is often impossible for computer users to browse documents that they are interested in one by one. Automatic document summarization involves condensing a document or documents to produce a fair summary with bearable time cost. Usually two kinds of summarization approaches can be employed: abstractive or extractive. Abstractive approaches typically need "understand" contents of documents. As the limitation of current NLP techniques, research on abstraction is confined in limited domains or in stage of preliminary investigations. Extractive summarization is investigated in this study with attempts to balance satisfactory performance and reasonable cost.

There is a common architecture for extractive summarization, which includes five sequential components: sentence (or paragraph) segmentation, weighting, ranking, re-ranking and summary generation. Given document(s), they should be segmented first. Then sentences (or paragraphs) will be weighted according to certain strategy. Later on they will be ranked based their weight scores. Sometimes the importance of sentences (or paragraphs) needs to be adjusted (e.g, to avoid redundancy) and they should be re-ranked based on a predefined algorithm. Finally important sentences will be extracted and included in final summaries. The core of this architecture is the sentence (or paragraphs) weighting component and kinds of approaches are proposed in the past.

Previous extractive summarization approaches identify the important content mainly based on terms. Bag-of-words is not a good representation to specify contents of documents. There are many possible combinations for the same collection of words. For example, given a sentence “Cat eat mouse”, the word set is {cat, eat, mouse}. The possible word combinations include “cat eat mouse”, “cat mouse eat”, “eat cat mouse”, “eat mouse cat”, “mouse eat cat” and “mouse cat eat”. Therefore, there are multiple different explanations for this word set. A predefined template is a better choice to represent the contents. For example, the template {subject: cat; object: mouse; action: eat} has only one exact explanation. However template is domain-dependent and need much effort to create and fill it. This tension motivates us to seek a balance between effective implementation and deep understanding, such as an automatically built semi-structure. It should be easier to achieve than pre-defined templates, and is more exact than bag-of-words.

According to [Filatovia and Hatzivassiloglou 2004], event may be a natural unit to convey meanings of documents. It can be broadly defined as “who did what to whom when and where”. Both linguistic and empirical studies acknowledge that event arguments help characterize the effects of a verb’s event structure, even though verbs or other words denoting event determine the semantics of an event. In this study, event is defined as event term and associated event elements. We choose verbs (such as “elect” and “incorporate”) and action nouns (such as “election” and “incorporation”) as event terms that can characterize actions. They roughly relate to “did what”. One or more associated named entities are considered as event elements. Four types of named entities are currently under consideration. They are person names, organization names, location and time. They convey the information about “who”, “whom”, “where” and “when”. To filter pseudo event terms, a verb or an action noun is deemed as an event term only when it occurs at least once between two named entities.

We investigate two different event representations in this study. One occurrence of an event term (or element) in a document is an instance of the event term (or element), while

the collection of the same event term (or element) is a concept of the event term (or element). For convenience, we do not discriminate instance of event term and instance of event element, all of them are called under the name “instance of event”. Similarly, concept of event term and concept of event element are called under the name “concept of event”. Therefore, contents of documents can be represented based on instances or concepts of events.

It is most likely that documents narrate more than one related event. Motivated by this observation, we investigate relevant event-based summarization after the study on independent event-based summarization. In our preliminary summarization experiments based on event instances, it is found that the performance is improved when relevance is considered. Then we conduct investigations on further experiments based on event concepts with relevance. Different approaches are designed to evaluate the strength of relevance.

The organization of Chapter 3 is described as follows: Section 3.2 describe the common architecture for extractive summarization. Section 3.3, Section 3.4 and Section 3.5 present summarization approaches based on independent event instances, relevant event instances and relevant event concepts, respectively. Section 3.6 describes event-based summarization approaches on different granularities. Section 3.7 presents the experiments of these different event-based extractive summarization approaches on DUC 2001 data set-- an English news document set. Section 3.8 analyzes and discusses the experiments and Section 3.9 summarizes this chapter.

3.2 Architecture for Extractive Summarization

Since extractive summarization has no need to deeply represent and understand contents of documents, it is a relatively applicable solution. It has received much more attention than ever. Kinds of extractive approaches have been proposed in the past. However, a common architecture is employed by these extractive approaches. It is described Figure 3.1.

Document(s) to be summarized should be segmented first. If the unit for extraction is paragraph, then documents will be divided into paragraphs. If the unit is sentence, then documents will be divided into sentences. In this study, extractive summarization with fine granularity unit -- sentence is investigated. Here sentence boundary identification is a problem to be addressed, as sometimes the punctuation mark "." does not denote ending of a sentence. For example, "Mr. Smith went to another city yesterday." Multiple tool packages have been implemented to identify sentence boundaries and they can be integrated into extractive summarization systems.

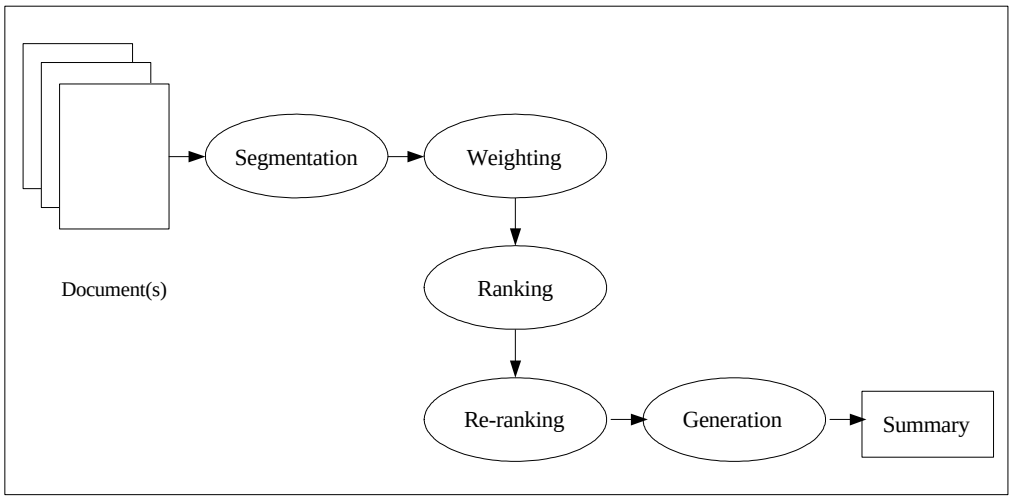

Figure 3.1. A common architecture for extractive summarization

Next, importance of sentences (or paragraphs) should be weighted according to certain strategies. Typical features which are used to evaluate the importance include the number of high frequency words or cue phrases, location of sentences and length of sentences. It is difficult to evaluate the importance of a sentence or paragraph based only one feature, therefore multiple features are employed. A linear function or a learning classifier is used to integrate these features and assign a weight score for each sentence or paragraph. This component is the core of the whole summarization architecture. Multiple weighting strategies, such as event-based, temporal-oriented event-based and learning-based ones are investigated in this study. In this chapter, sentences are weighted based on event instances

and event concepts. Section 3.3, Section 3.4 and Section 3.5 will describe summarization approaches that employ these weighting strategies.

Then sentences (or paragraphs) are ranked according to their weight scores. The most important sentences (or paragraphs) will be selected to generate summaries. However, sometimes a re-ranking algorithm is necessary to adjust the weight scores of sentences (or paragraphs). For example, the importance of redundant sentences can be deducted, so that same or similar sentences are not included into summaries. If the importance of sentences (paragraphs) is 0 or 1, and there are too many sentences (paragraphs) whose weight scores are 1, a re-ranking algorithm can be used to discriminate the degree of importance. The final weight scores are given after the re-ranking procedure. The re-ranking component is omitted in summarization approaches in this chapter, but it is employed in those in Chapter 6.

At last importance sentences (or paragraphs) are extracted and included into summaries until the length limitation of summary is reached. To form a natural summary, it is better to organize these important sentences (or paragraphs) according to certain order, such as the order that sentences (or paragraphs) occur in documents or the order that sentences (or paragraphs) occur on the time line. Since the focus of this study is to identify important contents of documents, this ordering procedure is omitted. Important sentences are extracted one by one according to their weight scores.

3.3 Independent Instance-based Summarization

Sentences are weighted by independent event instance based approaches in Section 3.3. As the fact, same or similar events may be mentioned for multiple times in documents. Events under different contexts may have different importance for representing the focus of the documents. It is found that important event terms are repeated and always occur with more named entities, because authors hope to state these events clearly. At the same time, people may want to emphasize an event by presenting it with action, time, location and

participants completely. They may omit time, location and participants of an event after they describe the event previously, or they just want to give a rough description.

It is assumed that events in documents may have different importance. In this study, event terms occurring in different circumstances are assigned different weights. Event terms between two named entities may be more important than event terms just beside one named entity. Event terms co-occurring with participants may be more important than event terms just beside time or location.

The approach on independent event-based summarization involves the following steps.

1. Given a cluster of documents, analyze sentences one by one. Named entities and POS tags are identified by GATE [Cunningham et al. 2002]. Ignore sentences that do not contain any event element.
2. Add a frequent noun into the set of named entities (NE) when the number of its occurrences is above a certain threshold.
3. After stemming each word, detect pairs of named entities in every sentence and extract verbs and action nouns as event terms (ET) Stop words are ignored in this procedure.
4. Scan documents again to extract events, which are event terms with adjacent named entities. An event takes the form as triple $(et_x \mid ne_i, ne_j)$, if the event term is between a pair of named entities; or as couple $(et_y \mid ne_k)$, if the event term is neighboring with only one named entity in a sentence.
5. Assign different weights to different event terms, according to contexts of event terms. Different weight configurations are described in Section 3.7. Contexts refer to number of event elements beside event terms and types of these event elements.
6. Given a cluster of multiple documents about a topic, for instances of same event term (or element) in one document, we calculate an average tf*idf score. There are multiple tf*idf scores for an event term (or element) in the cluster of documents and the average score is the weight of the event term (or element) in the cluster. The algorithm

is similar with Centroid, but in that work the weight is calculated based on word, not event term or event element.

7. Sum up the weights of event terms and event elements in a sentence.
8. Select the top sentences with higher weight scores according to the length limitation of summary.

3.4 Relevant Instance-based Summarization

The previous independent event-based approach does not exploit relevance between event instances. However, the relevance may be useful to identify important event instances. After a document is represented by event instances, relevant instances are linked together. Therefore the contents of a cluster of documents can be represented by an event map. It is assumed that important events may be mentioned often and events associated to important events may be important also. Based on this assumption, PageRank [Page et al. 1998] is a suitable algorithm to identify the importance of event instances in the map. In this section, we will discuss how to represent documents by event instances and how to identify important event instances with PageRank algorithm.

3.4.1 Document Representation by Instances

Events are commonly related with one another semantically, temporally or spatially, especially when the documents to be summarized are about the same or similar topics. Therefore, all event terms and named entities involved can be explicitly connected or implicitly related and weave a document or a set of documents into an event map. The procedure of event identification is same with that in independent instance-based summarization. Different instances which are about the same concept will be kept and linked in the map.

Given a sentence “Andrew had become little more than a strong rainstorm early yesterday, moving across Mississippi State and heading for the north-eastern US”, the event

map is shown in Figure 3.2. There are two kinds of nodes in the graph. Event terms (ET) are indicated by rectangles and named entities (NE) are indicated by ellipses. They represent instances of events. The links between every two nodes are undirectional.

After each sentence is represented by a map, there will be multiple sub-maps for a cluster of documents. If nodes from different sub-maps are lexical match, they may denote same thing and should be linked. For example, if named entity "Andrew" occurred in sentence *A*, *B* and *C*, then the three occurrences O_A, O_B and O_C will be linked as O_A—O_B, O_B—O_C, O_C—O_A. By this way, maps for sentences can be linked based on same concepts.

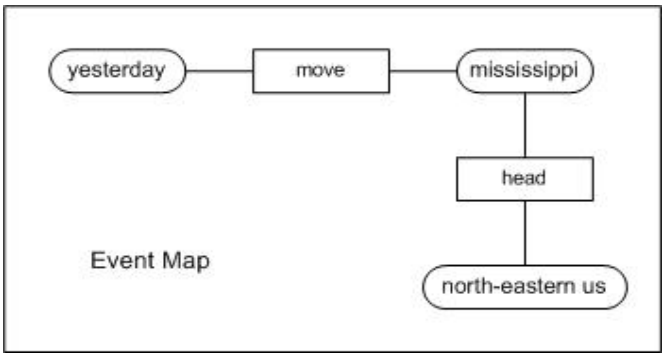

Figure 3.2 An example of instance-based event map

The advantage of representing with separated action and entity nodes over simply combining them into one event or sentence node is to provide a convenient way for analyzing the relevance among event terms and named entities either by their semantic or distributional similarity. More importantly, this favors extraction of concepts and brings the conceptual compression available. We call this representation an event map, from which the most important instances can be picked out for the summary.

3.4.2 Identification of Instance Importance

Given a whole map for a cluster of documents, the next step is to identify focus of these documents. Based on our assumption about important contents in the previous section, PageRank algorithm [Page et al. 1998] is employed to fulfill this task. PageRank assumes that if a node is connected with more other nodes, it may be more likely to be a salient

instance. The nodes relevant to the significant nodes are more likely salient instance than those not. The algorithm assigns the significance score to each node according to the number of nodes linking to it as well as the significance of the nodes. In PageRank algorithm, we use two directional links instead for every unidirectional link in Figure 3.2. For example, if a node *A* and another node *B* are connected in Figure 3.2, there will be two links between the node *A* and *B*: *A* → *B* and *B* → *A*.

The equation to calculate the importance (indicated by PR) of a certain node *A* is shown as follows:

$$PR(A) = (1-d) + d(\frac{PR(B_1)}{C(B_1)} + \frac{PR(B_2)}{C(B_2)} + ... + \frac{PR(B_t)}{C(B_t)})$$

Where B_1, B_2,…, B_t are all nodes which link to the node *A*. C(B_i) is the number of outgoing links from the node B_i. The weight score of each node can be achieved by this equation recursively. *d* is the factor used to avoid the limitation of loop in the map structure. As the literature [Page et al. 1998] suggested, *d* is set as 0.85. The significance of each sentence to be included in the summary is then derived from the significance of the event terms and named entities it contains.

3.5 Relevant Concept-Based Summarization

A collection of instances of same event term (or named entity) most likely represent a same concept. When people express their thinking by text, they mean concepts and the relations between concepts, not discrete instances of concepts. To simulate this procedure, we investigate event-based summarization based on concepts, i.e. collections of same instances. According to the preliminary experiments on instance-based summarization, it is found that relevance between events is helpful to improve the performance. Therefore we conduct our research on concept-based summarization with relevance only.

Event map can be constructed based on event concepts. The procedure to extract event instance is same as those of instance-based summarization approaches, which is presented in

Section 3.3. However, in concept-based summarization, if two same event instances are found after the procedure of event extraction, they will be merged as one node to construct the event map. The number of nodes in concept-based map is smaller than that in instance-based map. Similar with relevant instance-based map, links between nodes are unidirectional in concept-based map.

Intra-event and inter-event relevance are two kinds of relevance in the map. We study the relationship between event terms and associated named entities. We also study the relationship between event terms and event terms, or between named entities and named entities. These kinds of event relevance are both necessary to build a complete event map, which reflects relationships inside and between events. Section 3.5.2 and Section 3.5.3 describes intra-event and inter-event relevance respectively.

Given event map, Section 3.5.4 identifies the importance of nodes in the map with a link analysis algorithm -- PageRank. This procedure is similar with that in relevant instance-based approach. After the recursive computation of PageRank, weight of each event concept is achieved and it is used for every instance of this concept. Finally, the importance of each sentence is sum of all the weights of event instances contained in it.

3.5.1 Document Representation by Concepts

As discussed in previous sections, an event is defined as an event term and associated named entities. The event term should occur between two named entities at least once in documents. The procedure to form the event map is similar with that in the instance-based summarization, except that nodes should be concepts in this section. Words in either their original form or morphological variations are represented with only one node in the graph regardless of how many times they appear in documents. An example document is shown in Figure 3.3 and the corresponding concept-based event map is given in Figure 3.4.

We then integrate the strength of the connections between nodes into this graphical model in terms of the relevance defined from different perspectives. The relevance is

indicated by $r(node_i, node_j)$, where $node_i$ and $node_j$ represent two nodes, and are either event terms (et_i) or named entities (ne_j). Then, the significance of each node, indicated by $w(node_i)$, is calculated with PageRank ranking algorithm. Sections 3.5.2 and 3.5.3 address the issues of deriving $r(node_i, node_j)$ according to intra- or/and inter- event relevance.

S1: The Justice Department and the 20 states suing Microsoft believe that the tape will strengthen their case because it shows Gates saying he was not involved in plans to take what the government alleges were illegal steps to stifle competition in the Internet software market.

S2: It showed a few brief clips of a point in the deposition when Gates was asked about a meeting on June 21, 1995, at which, the government alleges, Microsoft offered to divide the browser market with Netscape and to make an investment in the company, which is its chief rival in that market.

S3: In the taped deposition, Gates says he recalled being asked by one of his subordinates whether he thought it made sense to invest in Netscape.

S4: But in an e-mail on May 31, 1995, Gates urged an alliance with Netscape.

S5: The contradiction between Gates' deposition and his e-mail, though, does not of itself speak to the issue of whether Microsoft made an illegal offer to Netscape.

Figure 3.3. A sample document

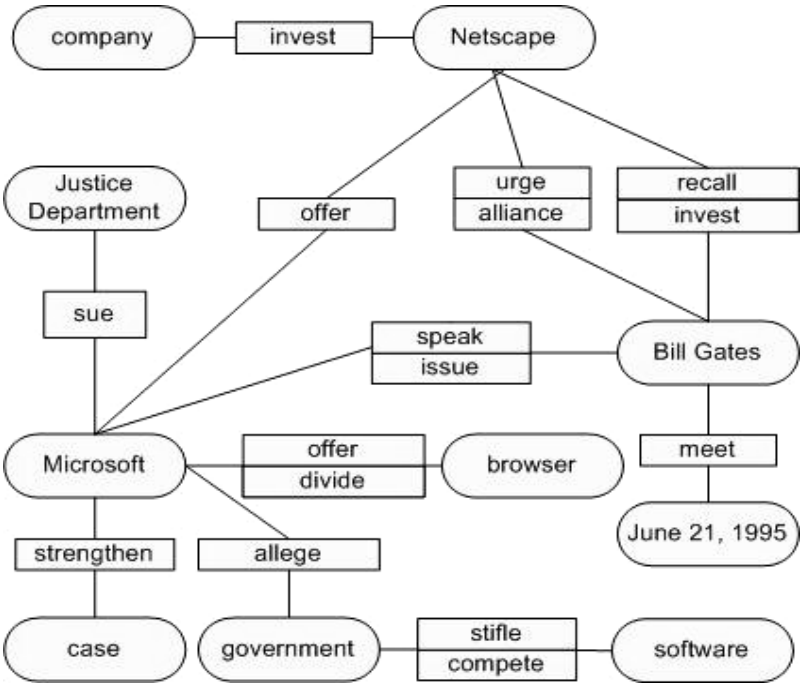

Figure 3.4. An example of concept-based event map

3.5.2 Intra-Event Relevance Weighting

We consider both intra-event and inter-event relevance for summarization. Intra-event relevance measures how an action itself is associated with its associated named entities. It is indicated as $R(ET, NE)$ and $R(NE, ET)$ in Table 3.1 below. This is a kind of direct relevance as the connections between actions and named entities are established from the text surface directly. No inference or background knowledge is required. As the connection between an event term et_i and a named entity ne_j is supposed symmetry, $R(NE, ET) = R(ET, NE)^T$. By means of inter-event relevance, we consider how an event term (or a named entity involved in an event) associate to another event term (or another named entity involved in the same or different events) from syntactic, semantic and distributional aspects. It is indicated by $R(ET, ET)$ or $R(NE, NE)$ in Table 3.1. Indirect connections which are not explicit in the event map should be derived from external resources or overall event distribution.

	Event Term (*ET*)	Named Entity (*NE*)
Event Term (*ET*)	$R(ET, ET)$	$R(ET, NE)$
Named Entity (*NE*)	$R(NE, ET)$	$R(NE, NE)$

Table 3.1 Relevance matrix

The complete relevance matrix is:

$$R = \begin{bmatrix} R(ET, ET) & R(ET, NE) \\ R(NE, ET) & R(NE, NE) \end{bmatrix}$$

The intra-event relevance $R(ET, NE)$ can be simply established by counting how many times et_i and ne_j are associated, i.e.

$$r_{Document}(et_i, ne_j) = freq(et_i, ne_j) \qquad \text{(E1)}$$

On the other hand, we observe in real texts that two named entities can be far apart in a long sentence and more than one event terms emerge between them (e.g. event terms “stifle” and “compete” in Figure 3.5; event terms in joined rectangles in Figure 3.4). These adjacent event terms which are associated with same pair of named entities are mostly because of

complicated sentence structure, such as subordinate clause. The weight of the connection between the same pair of named entities may be divided into each path via every different event term between the two named entities. The strength of relevance between an event term and a named entity within an event is indicated as $link(et_i, ne_j) = 1/n$, when n is the number of adjacent event terms between the same named entity (pair). The weight of connection between an event term and a named entity in the event map is calculated as the following equation. Figure 3.6 enlarges a sub-map of Figure 3.4 to show the weight of edges.

$$r_{split}(et_i, ne_j) = \sum link(et_i, ne_j) \qquad \text{(E2)}$$

Original:

The <Organization>Justice Department</Organization> and the 20 states <VB>suing</VB> <Organization>Microsoft</Organization> believe that the tape will <VB>strengthen</VB> their <HN>case</HN> because it shows <Person>Gates</Person> saying he was not <VB>involved</VB> in plans to take what the <HN>government</HN> alleges were illegal steps to <VB>stifle</VB> <AN>competition</AN> in the Internet <HN>software</HN> <HN>market</HN>.

Events:
1. {sue | Justice Department, Microsoft}
2. {strengthen | Microsoft, case}
3. {involve | Gates, government}
4, 5. {stifle, compete | government, software}

Figure 3.5 Events extracted from a sentence

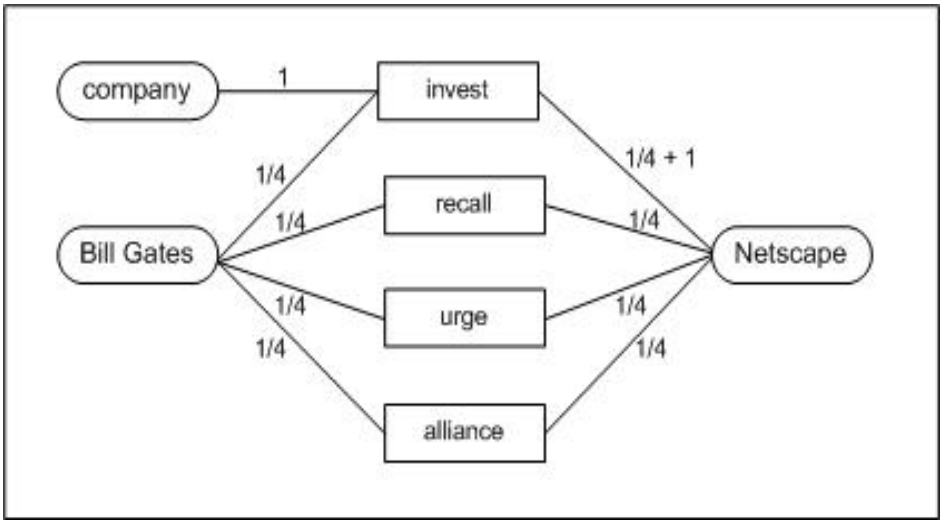

Figure 3.6 Weight of connections between event terms and named entities

3.5.3 Inter-Event Relevance Weighting

One way to measure the term relevance is to make use of a general language knowledge base, such as WordNet [Fellbaum 1998]. WordNet::Similarity is a freely available software package that makes it possible to measure the semantic relatedness between a pair of concepts, or in our case event terms, based on WordNet [Pedersen et al. 2004]. It supports three measures. The one we choose is the function *lesk*.

$$r_{WordNet}(et_i, et_j) = similarity(et_i, et_j) = lesk(et_i, et_j) \qquad \text{(E3)}$$

Alternatively, term relevance can be measured according to their distributions in the specified documents. We believe that if two events are concerned with the same participants, occur at same location, or at the same time, these two events are interrelated with each other in some ways. This observation motivates us to derive event term relevance from the number of name entities they share.

$$r_{Document}(et_i, et_j) = |\, NE(et_i) \cap NE(et_j) \,| \qquad \text{(E4)}$$

$NE(et_i)$ is the set of named entities which are associated with et_i. The symbol "| |" indicates the number of the elements in the set. The relevance of named entities can be derived in a similar way.

$$r_{Document}(ne_i, ne_j) = |\, ET(ne_i) \cap ET(ne_j) \,| \qquad \text{(E5)}$$

The relevance derived with (E4) and (E5) are indirect relevance. In previous work, a clustering algorithm, shown in Figure 3.7, has been proposed [Xu et al 2006] to merge the named entities that refer to the same concept (such as Ranariddh, Prince Norodom Ranariddh and President Prince Norodom Ranariddh). It is used for co-reference resolution and aims at joining instances with same concept into a single node in the event map. The experimental result suggests that merging named entities improves performance in some extent but not evidently. When applying the same algorithm for clustering all four types of name entities in our experiments, we observe that named entities in the same cluster do not always refer to the same objects, even if they are indeed related in some way. For example, "Mississippi" is

a state in United States, while "Mississippi River" is the second-longest river in the United States and flows through "Mississippi".

Step1: Each name entity is represented by $ne_i = w_{i1}w_{i2}...w_{ik}$, where w_i is the ith word in it. The cluster it belongs to, indicated by $C(ne_i)$, is initialled by $w_{i1}w_{i2}...w_{ik}$ itself.

Step2: For each name entity

$ne_i = w_{i1}w_{i2}...w_{ik}$

For each name entity $ne_j = w_{j1}w_{j2}...w_{jl}$, if $C(ne_i)$ is a sub-string of $C(ne_j)$, then $C(ne_i) = C(ne_j)$.

Continue Step 2 until no change occurs.

Figure 3.7 The algorithm proposed to merge the named entities

Location	Person	Date	Organization
Mississippi	Professor Sir Richard Southwood	first six months of last year	Long Beach City Council
Mississippi River	Sir Richard Southwood	last year	San Jose City Council
	Richard Southwood		City Council

Table 3.2 Some results of the named entity merged

It therefore provides a second way to measure named entity relevance based on the clusters found. It is actually a kind of measure of lexical similarity.

$$r_{Cluster}(ne_i, ne_j) = \begin{cases} 1, & ne_i, ne_j \text{ are in the same cluster} \\ 0, & \text{otherwise} \end{cases} \quad \text{(E6)}$$

In addition, the relevance of the named entities can be sometimes revealed by sentence context. Take the following most frequently used sentence patterns as examples:

<Person>, a-position-name of <Organization>, does something.
<Person> and another <Person> do something.

Considering that two neighboring name entities in a sentence are usually relevant, the following window-based relevance is also experimented with.

$$r_{pattern}(ne_i, ne_j) = \begin{cases} 1, & ne_i, ne_j \text{ are within a pre-specified window size} \\ 0, & \text{otherwise} \end{cases} \quad \text{(E7)}$$

Besides the previous methods to evaluate the inter-event relationships, there is another new point of view. The connection between two event terms (or two named entities) can be regarded as two continuous links: from one event term (or one named entity) to a named entity (or an event term), and then from the named entity (the event term) to another event term (or another named entity). Therefore the strength of the connection between the two event terms (or two named entities) can be represented by the production of the strength of the two links, i.e. two intra-event relationships. Based on this idea, the following two equations are given.

$$r_{bridge}(et_i, et_j) = \sqrt{\sum_{ne_x \in NE(et_i) \cap NE(et_j)} r_{split}(et_i, ne_x) \times \sum_{ne_x \in NE(et_i) \cap NE(et_j)} r_{split}(ne_x, et_j)} \quad \text{(E8)}$$

$$r_{bridge}(ne_i, ne_j) = \sqrt{\sum_{et_x \in ET(ne_i) \cap ET(ne_j)} r_{split}(ne_i, et_x) \times \sum_{et_x \in ET(ne_i) \cap ET(ne_j)} r_{split}(et_x, ne_j)} \quad \text{(E9)}$$

3.5.4 Identification of Concept Importance

The significance score, i.e. the weight $w(node_i)$ of each $node_i$, is then estimated recursively with PageRank ranking algorithm which assigns the significance score to each node according to the number of nodes connecting to it as well as the strength of their connections. The equation calculating $w(node_i)$ is shown as follows.

$$w(node_i) = (1-d) + d\left(\frac{w(node_1)}{r(node_i, node_1)} + \ldots + \frac{w(node_j)}{r(node_i, node_j)} + \ldots + \frac{w(node_t)}{r(node_i, node_t)}\right) \quad \text{(E10)}$$

In (E10), $node_j$ ($j = 1,2,\ldots t$, $j \neq i$) are the nodes linking to $node_i$. d is the factor used to avoid the limitation of loop in the map structure. It is set to 0.85 experimentally. It is assumed that instances that belong to the same concept have a same weight score. The significance of each sentence to be included in the summary is obtained from the significance of the event instances it contains. The sentences with higher significance are picked up into the summary as long as they are not exactly the same sentences. We are aware

of the important roles of information fusion and sentence compression in summary generation. However, the focus of this study is to extract the most important sentences and conceptual extraction is the future direction.

3.6 Summarization on Different Granularities

The contents of documents can be expressed with units of different granularities, such as event term / named entity, (event) and (sentence). To investigate the role of granularity, we compare results of summarization approaches on different granularities. In our experiments, we find that the performance of concept-based summarization is better than that of instance-based summarization. Therefore, we start this investigation with concept-based summarization. Concept-based summarization based on event term / named entity is presented in Section 3.5. Summarization approaches with the granularities of event and sentence are discussed in this section. In these two approaches, contents of documents are represented by instances of events or sentences, but the links between instances of events or sentences are measured by the strength of the links between concepts of event terms / named entities which are contained in the instances of events or sentences. Document representations with the granularities of event and sentence are shown in Figure 3.8 and Figure 3.9 respectively. Figure 3.9 can be derived from Figure 3.8, if instances of events which are contained in the same sentence are merged.

To construct the event map based on the granularity of event or sentence, events from each sentence are extracted. This procedure is same as those employed in event-based summarization approaches which are described in previous sections. Each event or each sentence will be a node in the map. The relevance between every two events or every two sentences is derived by sum all the weights of connections between event term and named entities, or similarly by sum all the weights of connections between events.

To determine the strength of events, there are two choices. One is to use a simple cosine similarity based on lexical event overlap and, the other is to use the strength of relations

between event elements. In this study, we focus on events not words, thus the second approach is more suitable to make use of event relevancy. As shown in Figure 3.8 and Figure 3.9, relations of events are measured by sum all the weights of connections between concepts of event terms and named entities, or similarly relations of sentence by weights of connections between events.

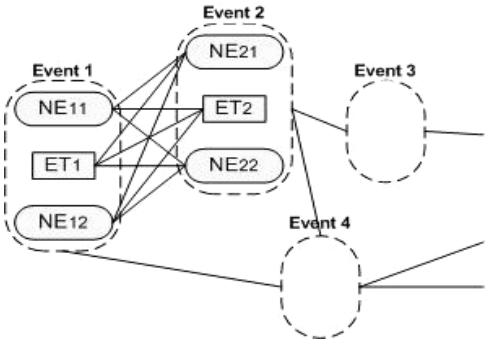

Figure 3.8 Map based on event

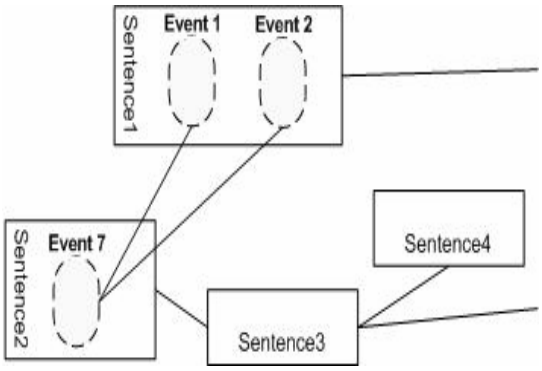

Figure 3.9 Map based on sentence

In summarization with the granularity of event, PageRank algorithm is employed to assign weight for each event. Then the importance of each sentence is achieved by sum weights of all instances of events contained in the sentence. In summarization with the granularity of sentence, then importance of sentence is identified by PageRank directly.

3.7 Experiment and Evaluation

3.7.1 Dataset and Evaluation Methods

To evaluate the event based summarization approaches proposed, we conduct a set of experiments on DUC 2001 dataset. Document Understanding Conferences (DUC) is a series of evaluation tasks on text summarization. They provide a public test benchmark for summarization systems. It contains 30 clusters of documents and a total of 308 English news reports. The number of documents in each cluster is between 3 and 20. The contents of each cluster are about certain news topic, such as the hurricane in Florida. An original document in this data set is shown in Figure 3.10.

<DOC>
<DOCNO>FT923-5835</DOCNO>
<PROFILE>_AN-CIBBCABPFT</PROFILE>
<DATE>920828
</DATE>
<HEADLINE>
FT 28 AUG 92 / UK Company News: GA says hurricane claims could reach 'up to Dollars 40m'
</HEADLINE>
<BYLINE>
By ROBERT PESTON
</BYLINE>
<TEXT>
GENERAL ACCIDENT, the leading British insurer, said yesterday that insurance claims arising from Hurricane Andrew could 'cost it as much as Dollars 40m.'
Lord Airlie, the chairman who was addressing an extraordinary shareholders' meeting, said: 'On the basis of emerging information, General Accident advise that the losses to their US operations arising from Hurricane Andrew, which struck Florida and Louisiana, might in total reach the level at which external catastrophe reinsurance covers would become exposed'.
What this means is that GA is able to pass on its losses to external reinsurers once a certain claims threshold has been breached.
It believes this threshold may be breached in respect of Hurricane Andrew claims.
However, if this happens, it would suffer a post-tax loss of Dollars 40m (Pounds 20m).
Mr Nelson Robertson, GA's chief general manager, explained later that the company has a 1/2 per cent share of the Florida market. It has a branch in Orlando.
The company's loss adjusters are in the area trying to estimate the losses.
Their guess is that losses to be faced by all insurers may total more than Dollars 8bn.
Not all damaged property in the area is insured and there have been estimates that the storm caused more than Dollars 20bn of damage.
However, other insurers have estimated that losses could be as low as Dollars 1bn in total.
Mr Robertson said: 'No one knows at this time what the exact loss is'.
</TEXT>
<PUB>The Financial Times
</PUB>
<PAGE>London Page 16
</PAGE>
</DOC>

Figure 3.10 A document in DUC 2001 data set

For each cluster, there are 3 different model summaries which contain same number of words. These summaries are provided manually. Three model summaries with 200 words for cluster d04 of DUC 2001 data set are shown in Figure 3.11 and Figure 3.12. These model summaries are created by NIST assessors for the DUC task of generic summarization. Manual summaries with 50 words, 100 words and 400 words are also provided. They are shown in Figure 3.12 and Figure 3.13.

Summary 1

Hurricane Andrew slammed across southern Florida and then continued into Louisiana where damage was limited because the storm narrowly missed low-lying New Orleans and the major oil industry complexes along the coast and offshore.

In Florida, wind gusts up to 165 mph left a ten-mile wide swath of death and destruction about 25 miles south of Miami.

The town of Homestead, including a local air force base, was largely flattened.

Total storm damage may exceed $20 billion, most of it not insured.

US property-casualty insurers expect to pay an estimated $7.3 billion in Florida alone.

This year, counting $3.9 billion for the LA riots and a series of tornadoes, plus Florida's Hurricane losses, the industry total rises to $11.2 billion.

This far exceeds the record $7.6 billion paid in 1989 for Hurricane Hugo and California's Loma Prieta earthquake.

Because US property- casualty insurers recently have purchased less foreign reinsurance for catastrophes, they alone will incur most of Andrew's losses.

Authorities believe they have adequate funds to cover the disaster.

After a poor initial response, President Bush seems finally to be getting disaster aid to those in need.

He has pledged to rebuild Homestead AFB and has sent the military to Florida.

Summary 2

Hurricane Andrew, the costliest natural disaster in US history, killed at least 17 people.

Southern Florida, in particular, Dade County was the scene of greatest damage.

One in every eight homes was destroyed.

In Florida overall, 150,000 persons were left homeless, and a week after the storm, 275,000 homes and businesses were still without electricity.

Louisiana was also severely damaged by Andrew.

It was initially feared that the storm might hit New Orleans which, because it is below sea level would be especially vulnerable.

However, Andrew made landfall 60 miles to the west and most of the extensive damage was to rural areas with the oil refining industry left mostly untouched.

US insurers expected Andrew claims could reach $8B.

Claims against British companies could reach $1B.

Total losses could be $15B with much of the damage to uninsured homes and businesses.

On-site officials in Florida were critical of delays in getting food, drinking water, and other needed supplies to the area.

Federal officials admitted problems and President Bush ordered troops to the area.

The Federal Emergency Management Agency, saddled with many political appointees, had no plan to deal with the disaster.

President Bush made a second trip to Florida and promised to rebuild Homestead Air Base.

Figure 3.11 Summaries with 200 words

Summary with 50 words

Damage in South Florida from Hurricane Andrew in August 1992 cost the insurance industry about $8 billion making it the most costly disaster in the US up to that time.

There were fifteen deaths and in Dade County alone 250,000 were left homeless.

Summary with 100 words

Hurricane Andrew which hit the Florida coast south of Miami in late August 1992 was at the time the most expensive disaster in US history.

Andrew's damage in Florida cost the insurance industry about $8 billion.

There were fifteen deaths, severe property damage, 1.2 million homes were left without electricity, and in Dade county alone 250,000 were left homeless.

Early efforts at relief were marked by wrangling between state and federal officials and frustrating delays, but the White House soon stepped in, dispatching troops to the area and committing the federal government to rebuilding and funding an effective relief effort.

Summary with 200 words

In late August 1992 Hurricane Andrew hit the Florida coast south of Miami with winds up to 165 mph, causing at least fifteen deaths, severe property damage and leaving 1.2 million homes without electricity.

In Dade County alone 250,000 were left homeless.

The town of Homestead and its nearby air force base were leveled.

As the storm continued across the Gulf of Mexico it was feared that it might hit New Orleans, but fortunately it made landfall in a relatively lightly populated area of Louisiana and quickly lost force as it moved over land.

The Federal Emergency Management Agency which was charged with handling such disasters had become a Republican patronage reserve and proved completely incapable of doing its job.

After frustrating delays in providing any effective relief, two of President Bush's assistants, Transportation Secretary Andrew Card and Chief of Staff James Baker, stepped in to do the job.

The president dispatched troops to Florida and gave a nationwide television address from the Oval Office promising to rebuild Homestead Air Force Base and committing the government to paying relief costs.

The insurance costs of Andrew alone were about $8 billion making it the most expensive disaster to have hit the US up to that time.

Figure 3.12 Summaries with 50, 100 and 200 words

Summary with 400 words

In late August 1992 Hurricane Andrew moved from the Bahamas across southern Florida and across the Gulf of Mexico into Louisiana.

As the hurricane hit the east coast of Florida south of Miami with winds gusting up to 165 mph, it ripped off roofs, smashed cars and trucks, snapped power lines, and uprooted trees.

There were at least fifteen deaths, severe property damage, and 1.2 million homes were left without electricity.

The storm destroyed the homes of one- eighth of the residents of Dade County leaving approximately 250,000 homeless.

The town of Homestead and its nearby Air Force Base were leveled.

As the storm continued across the Gulf of Mexico there was concern that it might hit New Orleans.

New Orleans, with a population of 1.6 million, is particularly susceptible to flooding since it lies below sea level, is intersected by the Mississippi River and has a large lake immediately to its north.

Officials in Louisiana, Mississippi, and Texas urged more than two million people to evacuate coastal areas.

Fortunately the hurricane missed New Orleans and made landfall in a relatively lightly populated area.

As it moved inland Andrew quickly lost force so that it was soon downgraded to tropical storm with winds at less than 75 mph.

Local officials in Florida were critical of a delay in supplying food, drinking water and other supplies for thousands of people in need.

The Federal Emergency Management Agency (FEMA) set up by President Carter in 1979 to handle disasters such as Andrew had become, under Republican administrations, the ultimate patronage backwater, having ten times as many political appointees as a typical agency.

FEMA was caught completely unprepared by Andrew.

There were unseemly disputes between state and federal authorities over who should do what in bringing relief.

Finally two of the President's right-hand men, Transportation Secretary Andrew Card and Chief of Staff James Baker, stepped in to do what FEMA should have done.

Bush ordered military forces to Florida to organize and run the relief effort and gave a televised speech to the nation from the Oval Office promising to rebuild Homestead Air Force Base and committing the government to paying the emergency relief costs.

Estimates of the cost of Hurricane Andrew vary but the insurance costs alone came to about $8 billion making it clearly the most expensive disaster ever to have hit the US up to that time.

Figure 3.13 A summary with 400 words.

In the following experiments, the DUC documents are segmented by a sentence boundary identifier. An example of segmented documents is shown in Figure 3.14. Then these segmented documents are tagged by GATE to recognize POS tags and four types of name entities. An example of tagged documents with POS tags and named entities are shown in Figure 3.15 and Figure 3.16 respectively. In our experiments different kinds of documents are kept separately. On average, each cluster of documents contains 10.3 documents, 602 sentences, 216 event terms and 148.5 name entities.

To evaluate the quality of the generated summaries, we choose an automatic summary evaluation package ROUGE [Lin and Hovy 2003], which has been used in DUC. ROUGE compares the machine generated summaries with manually provided summaries, based on unigram overlap and bigram overlap. A summarization approach will receive three evaluation scores from ROUGE: ROUGE-1 (unigram-based), ROUGE-2 (bigram-based), and ROUGE-W (based on longest common subsequence weighed by the length).

GENERAL ACCIDENT, the leading British insurer, said yesterday that insurance claims arising from Hurricane Andrew could 'cost it as much as Dollars 40m.'

Lord Airlie, the chairman who was addressing an extraordinary shareholders' meeting, said: 'On the basis of emerging information, General Accident advise that the losses to their US operations arising from Hurricane Andrew, which struck Florida and Louisiana, might in total reach the level at which external catastrophe reinsurance covers would become exposed'.

What this means is that GA is able to pass on its losses to external reinsurers once a certain claims threshold has been breached.

It believes this threshold may be breached in respect of Hurricane Andrew claims.

However, if this happens, it would suffer a post-tax loss of Dollars 40m (Pounds 20m).

Mr Nelson Robertson, GA's chief general manager, explained later that the company has a 1/2 per cent share of the Florida market.

It has a branch in Orlando.

The company's loss adjusters are in the area trying to estimate the losses.

Their guess is that losses to be faced by all insurers may total more than Dollars 8bn.

Not all damaged property in the area is insured and there have been estimates that the storm caused more than Dollars 20bn of damage.

However, other insurers have estimated that losses could be as low as Dollars 1bn in total.

Mr Robertson said: 'No one knows at this time what the exact loss is'.

Figure 3.14 An example of documents with segmented sentences

GENERAL/NNP ACCIDENT/NNP ,/, the/DT leading/VBG British/JJ insurer/NN ,/, said/VBD yesterday/NN that/IN insurance/NN claims/NNS arising/VBG from/IN Hurricane/NNP Andrew/NNP could/MD '/POS cost/VB it/PRP as/RB much/RB as/IN Dollars/NNS 40m/RB ./. '/"

Lord/NNP Airlie/NNP ,/, the/DT chairman/NN who/WP was/AUX addressing/VBG an/DT extraordinary/JJ shareholders/NNS '/POS meeting/NN ,/, said/VBD :/: '/" On/IN the/DT basis/NN of/IN emerging/VBG information/NN ,/, General/NNP Accident/NNP advise/VBP that/IN the/DT losses/NNS to/TO their/PRP$ US/NNP operations/NNS arising/VBG from/IN Hurricane/NNP Andrew/NNP ,/, which/WDT struck/VBD Florida/NNP and/CC Louisiana/NNP ,/, might/MD in/IN total/NN reach/VB the/DT level/NN at/IN which/WDT external/JJ catastrophe/NN reinsurance/NN covers/VBZ would/MD become/VB exposed/VBN '/POS ./.

What/WP this/DT means/VBZ is/AUX that/IN GA/NNP is/AUX able/JJ to/TO pass/VB on/RP its/PRP$ losses/NNS to/TO external/JJ reinsurers/NNS once/IN a/DT certain/JJ claims/NNS threshold/NN has/AUX been/AUX breached/VBN ./.

It/PRP believes/VBZ this/DT threshold/NN may/MD be/AUX breached/VBN in/IN respect/NN of/IN Hurricane/NNP Andrew/NNP claims/NNS ./.

However/RB ,/, if/IN this/DT happens/VBZ ,/, it/PRP would/MD suffer/VB a/DT post-tax/JJ loss/NN of/IN Dollars/NNS 40m/NNS -LRB-/-LRB- Pounds/NNS 20m/NNS -RRB-/-RRB- ./.

Mr/NNP Nelson/NNP Robertson/NNP ,/, GA/NNP 's/POS chief/JJ general/JJ manager/NN ,/, explained/VBD later/RB that/IN the/DT company/NN has/AUX a/DT 1/2/NN per/IN cent/NN share/NN of/IN the/DT Florida/NNP market/NN ./.

It/PRP has/AUX a/DT branch/NN in/IN Orlando/NNP ./.

The/DT company/NN 's/POS loss/NN adjusters/NNS are/AUX in/IN the/DT area/NN trying/VBG to/TO estimate/VB the/DT losses/NNS ./.

Their/PRP$ guess/NN is/AUX that/IN losses/NNS to/TO be/AUX faced/VBN by/IN all/DT insurers/NNS may/MD total/VB more/JJR than/IN Dollars/NNS 8bn/NN ./.

Not/RB all/PDT damaged/VBN property/NN in/IN the/DT area/NN is/AUX insured/VBN and/CC there/EX have/AUX been/AUX estimates/NNS that/WDT the/DT storm/NN caused/VBD more/JJR than/IN Dollars/NNS 20bn/NN of/IN damage/NN ./.

However/RB ,/, other/JJ insurers/NNS have/AUX estimated/VBN that/IN losses/NNS could/MD be/AUX as/RB low/JJ as/IN Dollars/NNS 1bn/VBN in/IN total/NN ./.

Mr/NNP Robertson/NNP said/VBD :/: '/POS No/DT one/NN knows/VBZ at/IN this/DT time/NN what/WP the/DT exact/JJ loss/NN is/AUX '/POS ./.

Figure 3.15 An example of documents with POS tags

GENERAL <Person>ACCIDENT</Person>, the leading British insurer, said <Date>yesterday</Date> that insurance claims arising from Hurricane <Person>Andrew</Person> could 'cost it as much as Dollars 40m.'

<Person>Lord Airlie</Person>, the chairman who was addressing an extraordinary shareholders' meeting, said: 'On the basis of emerging information, <Person>General Accident</Person> advise that the losses to their <Location>US</Location> operations arising from Hurricane <Person>Andrew</Person>, which struck <Location>Florida</Location> and <Location>Louisiana</Location>, might in total reach the level at which external catastrophe reinsurance covers would become exposed'.

What this means is that GA is able to pass on its losses to external reinsurers once a certain claims threshold has been breached.

It believes this threshold may be breached in respect of Hurricane <Person>Andrew</Person> claims.

However, if this happens, it would suffer a post-tax loss of Dollars 40m (Pounds 20m).

<Person>Mr Nelson Robertson</Person>, GA's chief general manager, explained later that the company has a 1/2 per cent share of the <Location>Florida</Location> market.

It has a branch in <Location>Orlando</Location>.

The company's loss adjusters are in the area trying to estimate the losses.

Their guess is that losses to be faced by all insurers may total more than Dollars 8bn.

Not all damaged property in the area is insured and there have been estimates that the storm caused more than Dollars 20bn of damage.

However, other insurers have estimated that losses could be as low as Dollars 1bn in total.

<Person>Mr Robertson</Person> said: 'No one knows at this time what the exact loss is'.

Figure 3.16 An example of documents with named entities

3.7.2 Experiments on Instance-based Event Summarization

In the following experiments for independent event-based summarization, we investigate the effectiveness of the approach. In addition, we attempt to test the importance of contextual information in scoring event terms. The number and type of neighboring named entities are considered to set the weights of event terms. The weight parameters in the following experiments are chosen according to empirical estimations.

Exp_AllSame: Weight of any named entity is 1. Weight of any verb/action noun, which is between two named entities or just beside one named entity, is 1.

Exp_EntNum: Weight of any named entity is 1. Weight of any verb/action noun, which is between two named entities, is 3. Weight of any verb/action noun, which is just beside one named entity, is 1.

Exp_EntType: Weight of any named entity is 1. Weight of any verb/action noun, which is between two named entities and the first named entity is person or organization, is 5. Weight of any verb/action noun, which is between two named entities and the first named entity is not person and not organization, is 3. Weight of any verb/action noun, which is just after a person or organization, is 2. Weight of any verb/action noun, which is just before one named entity, is 1. Weight of any verb/action noun, which is just after one named entity and the named entity is not person or organization, is 1.

In the following experiments, we investigate the effectiveness of our approaches under different length limitations of summaries. Based on the algorithm of Exp_EntType, we design experiments to generate summaries with length 50 words, 100 words, 200 words, 400 words. They are named **Exp_50**, **Experiment_100**, **Exp_200** and **Exp_400**.

In other experiments for relevant event-based summarization, we investigate the function of relevance between events. The configurations are described as follows.

Exp_Rel: Event terms and named entities are identified by the method we described in Section 3.3. In this experiment, frequent nouns are added to named entities. Occurrences of event terms or named entities are linked with by exact matches. Finally, the PageRank is employed to select important events and then important sentences are extracted.

Exp_Model: For reference, we select one of the three model summaries as the final summary for each cluster of documents. ROUGE is employed to evaluate the performance of these manual summaries.

The experiment results on independent event-based summarization are shown in table 3.3. The results for relevant event-based summarization are shown in table 3.5. From table 3.3, we can see that results of Exp_EntNum are similar with that of Exp_AllSame. It can be seen that importance of event terms is not very different when these event terms occur with

different number of named entities. Results of Exp_EntType are not significantly better than those of Exp_EntNum, so it seems that the importance of event terms is not very different when these event terms occur with different types of event elements. Therefore from these experiments, it can be seen that the number or type of named entities does not influence the importance of event terms significantly.

	Exp_AllSame	Exp_EntNum	Exp_EntType
ROUGE-1	0.315	0.316	0.318
ROUGE-2	0.049	0.051	0.054
ROUGE-W	0.110	0.110	0.111

Table 3.3 Evaluation results on independent instance-based summarization (summary with 200 words)

	Exp_50	Exp_100	Exp_200	Exp_400
ROUGE-1	0.182	0.243	0.318	0.386
ROUGE-2	0.019	0.031	0.055	0.080
ROUGE-W	0.091	0.108	0.124	0.135

Table 3.4 Evaluation results on independent event-based summarization (summary with different length)

Four experiments of table 3.4 show that the performance of independent event based summarization is getting better, when the length of summaries is increased. One possible reason is that independent event based approach prefers sentences with more event terms and named entities, therefore the preferred length of sentences is longer. While in a short summary, people always condense sentences from original documents, and use some new words to substitute original concepts in documents. Then the ROUGE score, which evaluates word overlap, is not good in the event-based approach. In contrast, if the length of summaries is increased, people will adopt detail event descriptions in original documents, so the performance of summarization is improved.

In table 3.5, it can be seen that the ROUGE scores of relevant event-based summarization (Exp_Rel) are better than those of independent approach (Exp_AllSame). In Exp_AllSame, the weights of event element and named entities are not discriminated. In

Exp_Rel, the weights of event element and named entities are not discriminated. It is fair to compare Exp_Rel with Exp_AllSame, while it's unfair to compare Exp_Rel with Exp_EntType. It looks like the relevance between nodes (event terms or named entities) help to improve the performance. However, performance of both dependent and independent event-based summarization need to be improved further, compared with that of Exp_Model.

	Exp_Rel	Exp_Model
ROUGE-1	0.334	0.595
ROUGE-2	0.061	0.394
ROUGE-W	0.129	0.268

Table 3.5 Evaluation results on relevant event-based summarization and a reference experiment (summary with 200 words)

To remove redundancy, same sentences which are all important can not be included in summaries. Cosine similarity score is employed to evaluate every two important sentences. If the cosine similarity is above a predefined threshold, the similar sentence will not be included. However, it is found that this cosine similarity can not improve the summarization performance. Therefore, we just remove same sentences in following experiments.

3.7.3 Experiments on Concept-based Event Summarization

We first evaluate the summaries generated based on $R(ET, NE)$ itself. It means inter-event relationships are not considered. Here $R(ET, NE)$ is established by counting how many times et_i and ne_j are associated (see E1). In the pre-evaluation experiments, it is observed that some high frequency nouns, such as "doctors" and "hospitals", by themselves are not marked by general NE taggers. However, they really refer to persons, organizations or locations. We compare the ROUGE scores of adding frequent nouns or not to the set of named entities in Table 3.6. A noun is considered as a frequent noun when its frequency is larger than 10. Roughly 5% improvement is achieved when high frequent nouns are taken

into the consideration. It supports our decision to regard high frequent nouns as entities and employ them in the following experiments.

$R(ET, NE)$	*NE* Without High Frequency Nouns	*NE* With High Frequency Nouns
ROUGE-1	0.333	0.349
ROUGE-2	0.063	0.072
ROUGE-W	0.130	0.135

Table 3.6 ROUGE scores using $R(ET, NE)$

Different inter-event relationships are verified in the following experiments. If there is no extra specification, intra-event relationships which are evaluated by (E1) are incorporated into event map. Table 3.7 below then presents the summarization results by using $R(ET, ET)$. It compares two relevance derivation approaches, $R_{WordNet}$ and $R_{Document}$. The topic-specific relevance derived from the documents to be summarized outperforms the general purpose Word-Net relevance by about 4%. This result is reasonable as WordNet may introduce the word relatedness which is not necessary in the topic-specific documents. When we examine the relevance matrix from the event term pairs with the highest relevance, we find that the pairs, like "abort" and "confirm", "vote" and confirm", do reflect semantics (antonymous) and associated (causal) relations to some degree.

$R(ET, ET)$	Semantic Relevance from Word-Net	Topic-Specific Relevance from Documents
ROUGE-1	0.329	0.342
ROUGE-2	0.057	0.069
ROUGE-W	0.120	0.133

Table 3.7 ROUGE scores using $R(ET, ET)$

$R(NE, NE)$	Relevance from Documents	Relevance from Clustering	Relevance from Window-based Context
ROUGE-1	0.352	0.336	0.345
ROUGE-2	0.071	0.073	0.075
ROUGE-W	0.136	0.131	0.135

Table 3.8 ROUGE scores using $R(NE, NE)$

Surprisingly, the best individual result is from document distributional similarity $R_{Document}(NE, NE)$ in Table 3.8. Looking more closely, it can be seen that compared to event terms, named entities are more representative to the documents in which they are included. In other words, event terms are more likely to be distributed around all the document sets, whereas named entities are more topic-specific and more likely cluster in a particular document set. Examples of high related named entities in relevance matrix are "Andrew" and "Florida", "Louisiana" and "Florida". Although their relevance is not as explicit as the same of event terms (their relevance is more contextual than semantic), we can still deduce that some events may happen in both Louisiana and Florida, or about Andrew in Florida. In addition, it also shows that the relevance we would have expected to be derived from patterns and clustering can also be discovered by $R_{Document}(NE, NE)$. The window size is set to 3.8 experimentally in window-based practice.

$R(NE, NE)$	50	100	200	400
ROUGE-1	0.224	0.286	0.352	0.416
ROUGE-2	0.034	0.055	0.071	0.103
ROUGE-W	0.102	0.116	0.136	0.139
$R(ET, NE)$	50	100	200	400
ROUGE-1	0.222	0.279	0.349	0.416
ROUGE-2	0.033	0.051	0.072	0.104
ROUGE-W	0.102	0.115	0.135	0.139
$R(ET, ET)$	50	100	200	400
ROUGE-1	0.206	0.269	0.342	0.412
ROUGE-2	0.023	0.046	0.069	0.103
ROUGE-W	0.092	0.111	0.133	0.137
$R(ET, NE) + R(ET, ET) + R(NE, NE)$	50	100	200	400
ROUGE-1	0.213	0.279	0.346	0.417
ROUGE-2	0.031	0.051	0.071	0.106
ROUGE-W	0.095	0.114	0.134	0.139

Table 3.9 ROUGE scores using complete *R* matrix and with different summary length

Next, we evaluate the integration of $R(ET, NE)$, $R(ET, ET)$ and $R(NE, NE)$ (see Table 3.9). As DUC 2001 provides 4 different summary sizes for evaluation, it satisfies our desire to test the sensibility of the proposed event-based summarization techniques to the length of summaries. While the previously presented results are evaluated on 200 word summaries, now the results in four different sizes, i.e. 50, 100, 200 and 400 words are evaluated. The experimental results show that the event-based approaches indeed prefer longer summaries. This is coincident with what we have hypothesized. For this set of experiments, we choose to integrate the best method from each individual evaluation presented previously. It appears that using the named entity relevance which is derived from the event terms gives the best ROUGE scores in almost all the summery sizes. Compared with the results provided in [Filatova and Hatzivassiloglou 2004] whose average ROUGE-1 score is below 0.3 on the same data set, the significant improvement is revealed.

As discussed in Section 3.5, the named entities in the same cluster may often be relevant but not always be co-referred. In the following set of experiments (Table 3.10), we evaluate different two ways to use the clustering results. One is to consider them as related as if they are in the same cluster and derive $R(NE, NE)$ with (E6). The other is to merge the entities in one cluster as one reprehensive named entity and then use it in $R(ET, NE)$ with (E1). The rationality of the former approach is validated.

	Clustering is used to derive *NE-NE*	Clustering is used to merge entities and then to derive *ET-NE*
ROUGE-1	0.341	0.330
ROUGE-2	0.067	0.062
ROUGE-W	0.132	0.128

Table 3.10 ROUGE scores with regard to how to use the clustering information

In the last set of experiments on relevant concept-based summarization, we consider the linguistic structure of sentences. The connection between two event terms (or two named entities) can be regarded as two continuous links: from one event term (or one named entity) to a named entity (or an event term), and then from the named entity (the event term) to

another event term (or another named entity). Therefore the strength of the connection between the two event terms (or two named entities) can be represented by the production of the strength of the two links.

The strength of intra-event relationships $R(ET, NE)$ is evaluated by (E2), while the strength of inter-event relationships $R(ET, ET)$ and $R(NE, NE)$ is evaluated by (E8) and (E9). The results are shown in Table 3.11. For comparison, the performance of the previous relevant concept-based summarization in Table 3.9, i.e. $R(ET, NE)$ + $R(ET, ET)$ + $R(NE, NE)$ (200 words), is listed here. As shown in Table 3.11, a slight improvement is achieved by the new splitting approach.

	Co-occurrence	Indirect relevance
ROUGE-1	0.346	0.353
ROUGE-2	0.071	0.072
ROUGE-W	0.134	0.129

Table 3.11 ROUGE scores using different methods to weight relations in event map

3.7.4 Experiments on Event Summarization with Different Granularities

As discussed in Section 3.6, document map can be constructed by choosing different kinds of nodes. Table 3.12 shows the results of summarization approaches on different granularities. The advantage of representing with separated event terms and named entities over simply combining them into event or sentence node is to provide a convenient and effective way for analyzing the relevance between conceptual information. At the same time, the map on event or sentence level helps people to observe and investigate documents more conveniently.

Granularity	*ET*/*NE*	Event	Sentence
ROUGE-1	0.352	0.333	0.340
ROUGE-2	0.071	0.059	0.066
ROUGE-W	0.136	0.121	0.124

Table 3.12 ROUGE scores according to event maps based on different granularities

3.8 Discussion

Event-based approaches are also employed in previous works. We evaluate our work in this context. As event-based approaches in this paper are similar with that of Filatovia and Hatzivassiloglou [2004], and the evaluation data set is the same one, our results are compared with theirs.

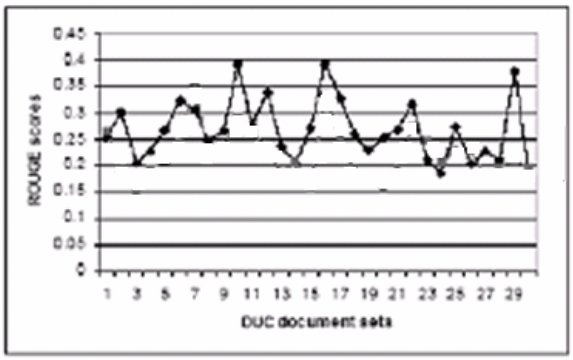

Figure3.17 Results of summarization reported in [Filatovia and Hatzivassiloglou 2004]

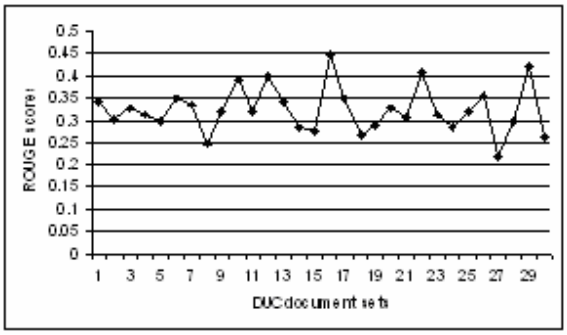

Figure 3.18 Results of our relevant instance-based summarization

Filatovia and Hatzivassiloglou [2004] report the ROUGE scores according to each cluster of DUC 2001 data collection in Figure 3.17. In this figure, the line represents their event-based approach. The evaluation of our relevant instance-based summarization approach presented in Section 3.4 is shown in Figure 3.18. The proposed approach achieves significant improvement on most document clusters. The reason seems that the relevance between events has been exploited.

Centroid-based summarization is one of successful term-based summarization approaches. It is a widely used and very challenging baseline in the text summarization

community. MEAD [Radev et.al. 2004] is employed to generate Centroid-based summaries. The ROUGE scores of Centroid-based summarization and those of relevant concept-based event summarization (Table 3.11 "indirect relevance") are compared. The scores are reported for each cluster of documents (Figure 3.19). Finally, for 18 clusters (60%) out of the 30 clusters of documents, the summary created according to document graph (i.e., event map) with frequent nouns counted in receives higher ROUGE score than according to Centroid-based approach. By taking high frequency nouns into the consideration, great improvement is achieved in 20 clusters (66.7%) and 5% increase of ROUGE score is gained on average. The advantage of graph-based approaches over Centroid-based one is that they indicate redundant information by link weight and prevent rare words with high *idf* scores which are unrelated to the topic.

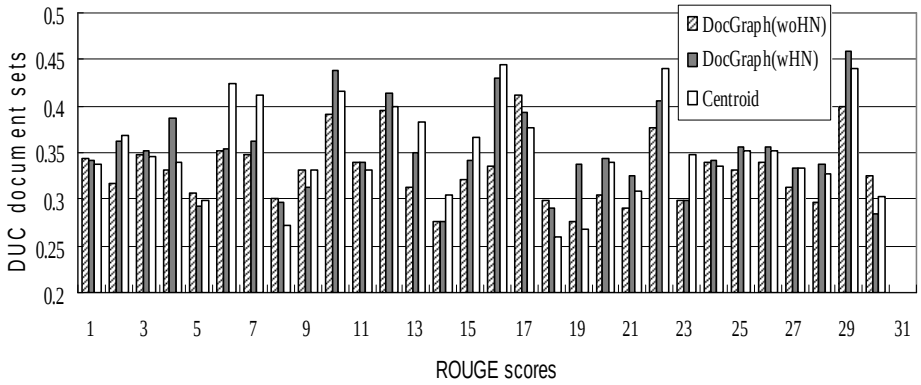

Figure 3.19 Results of summarization approaches based on document graph (with and without high frequency noun) or Centroid

3.9 Chapter Summary

In this study, we propose to integrate event-based approaches to extractive summarization. An event-based scheme is employed to represent document and identify important contents. The independent instance-based summarization identifies important contents according to event frequency. We also investigate the discrimination of event terms in different context. Experiments show that it is not very helpful to improve the final performance. Therefore we do not consider the influence of number and types of named

entities on the importance of event terms in the following study. Then we explore summarization under different length limitation. It can be seen that our independent instance-based summarization acts well with longer summaries. In the relevant instance-based summarization approach, events are linked together by same or similar event terms and named entities. Experiments show that the relevance between events can improve the performance of summarization. Compared with those of close related work, our approaches achieve encouraging improvement.

In relevant concept-based summarization, both inter-event and intra-event relevance are investigated. PageRank algorithm is used to evaluate the significance of each concept (including both event terms and named entities). The sentences containing more concepts and highest significance scores are chosen in the summary as long as they are not the same sentences. To derive event relevance, we consider the associations at the syntactic, semantic and contextual levels. An important finding on the DUC 2001 data set is that making use of named entity relevance derived from the event terms that they associate with achieves the best result. The ROUGE-1 score 0.352 significantly outperforms the one reported in the closely related work whose average is below 0.3.

We are interested in the issue of how to improve event representation in order to build a more powerful event-based summarization approach. This would be one of our future directions. We also want to see how concepts rather than sentences are selected into the summaries in order to develop a more flexible compression technique and to know what characteristics of a document cluster is appropriate for applying event-based summarization techniques.

Chapter 4

Temporal Expression Extraction and Normalization

4.1 Chapter Overview

Temporal information processing is valuable in many NLP applications, such as text summarization, information extraction, machine translation and question answering. In this study, i.e. temporal-oriented event-based summarization, temporal information is crucial to anchor events on the time line. However, a wide scope of linguistic means, from lexical to syntactic phenomena, can represent this information. It is hard to catch the internal temporal meanings which are behind surface texts with various forms. Furthermore, same text combined with different context may indicate different interpretations, either temporal or non-temporal.

Temporal expressions convey crucial temporal information for anchoring events. In this study, temporal expressions are defined as chunks of text which convey the knowledge about time point or duration. A time point refers to a region which can be anchored on the time line. Normally, the length of the region is just a certain temporal unit, such as century, day or minute. “July 15, 1999”, “Thursday”, “yesterday”, “1960s”, “Ten minutes to 3”, “twelve o’clock January 3, 1984” and “11:59 p.m.” are all time points. Duration indicates a period of time, i.e. how long something lasts, such as “two hours” and “the past four years”. If the period can be anchored on the time line, normally it can not be represented by the combination of an ordinal number and a temporal unit.

TIMEX2 annotating guidelines [Ferro et al. 2004; Gerber et al. 2004] give detail descriptions about temporal expressions. According to the guidelines, temporal expressions include date, time, duration, frequency, event-anchored expressions, and so on. To retrieve the useful temporal information contained in these temporal expressions, we need to identify the extents of temporal expressions in raw text and then represent temporal attributes of the expressions according to the standard. For example, "July 15, 1999" and "1960s" should be represented by "<TIMEX2 VAL="1999-07-15">" and "<TIMEX2 VAL= "196">". The two tasks are called temporal extraction and temporal normalization, respectively.

A variety of efforts have been devoted to temporal information processing, such as 2001 Annual Meeting of the Association for Computational Linguistics (ACL) workshop on temporal and spatial information processing and 2002 International Conference on Language Resources and Evaluation (LREC). Some temporal parsers have been developed to extract normalize temporal expressions, but the performance can be improved further or the cover scope of temporal expressions can be extended. It motivates us to develop an approach which can automatically identify various temporal expressions with high precision.

As the temporal processing model will be integrated with the summarization systems on English and Chinese texts, the two languages are under consideration in this chapter. The temporal expression extraction and normalization techniques on the two languages are similar, even if some difference exists, such as the problem of word segmentation in Chinese. We implement the temporal expression extraction and normalization approaches on Chinese texts first [Wu et al. 2005b]. Then we tune these approaches on English documents later. Two full systems according to TIMEX2 guidelines, ETEMP and CTEMP, have been implemented on English and Chinese documents respectively. Each of them consists of two modules: extractor and normalizer.

Although machine-learning based approach can be used for temporal expression extraction, they are not suitable for the task normalization [Wu et al. 2005a; Wu et al. 2006]. Effective features are difficult to achieve for the mapping from temporal expressions to

values of attributes. Rule based approach is a conventional approach for both extraction and normalization. As the scope of the problem to identify and normalize temporal expressions is rather limited, a deliberate rule set is competent. When we implement our two systems for English and Chinese texts, they are tuned on corresponding data set respectively.

To design the rules for extraction and normalization, we study basic temporal objects and relations, the measurement of time and the classification of temporal expressions. A chart parser is employed to recognize temporal expressions based on these rules. When the temporal expressions are recognized, their temporal attributes are interpreted instantly, i.e. fill values for corresponding temporal attributes. Experiments show that our rule-based approach achieves promising results no matter on English or Chinese texts. The performance of English Temporal system (ETEMP) is medium among those of other English systems when evaluated by TERN 2004 [TERN 2004]. Chinese Temporal (CTEMP) has been evaluated by TERN 2004, ACE 2005 and ACE 2007. The performance is ranked first in these evaluation tasks.

Chapter 4 is organized as follows. Section 4.2 presents English temporal expression extraction and normalization, while Section 4.3 presents Chinese temporal expression extraction and normalization. In each of the two sections, temporal expression classification, grammar rules, constraint rules, and normalization procedure are described in detail. Section 4.4 presents experimental configuration and evaluation results on English and Chinese. Finally, Section 4.5 summarizes this chapter.

4.2 English Temporal Expression Processing

4.2.1 English Temporal Expression Classification

According to our observation on news documents and the guideline of TERN 2004, temporal expressions can be classified into different classes (Figure 4.1.). Although these

temporal expressions are from news documents, they cover most common representations in English documents.

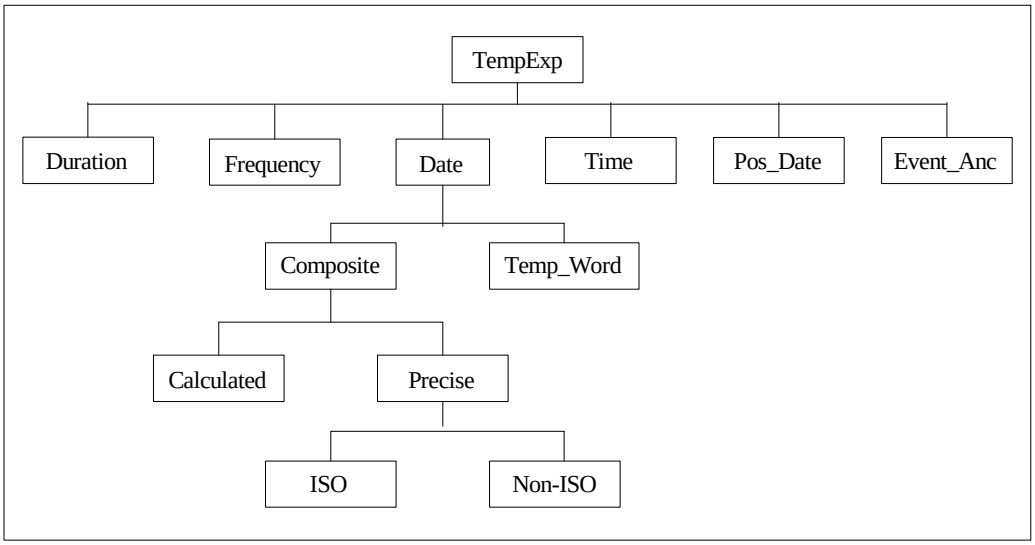

Figure 4.1 The classification of English temporal expressions

4.2.1.1 Time

These expressions are used to express a time value in some day, such as "15:30" and "half past three". If they are used together with date expressions, we can anchor the time value on the time line. There are three constraints in this procedure. Temporal expressions of the kind "Time" should lie behind the "Date" expression. The minimal unit level in "Date" expressions should be "day". The maximal unit level in "Time" expression should be "hour". Typical temporal expressions of the kind "Time" are listed as follows.

18:47:10.19	5:10 AM EST
1:46:44.47	1123EDT
4:30 p.m	0925EDT
0417 GMT	0611EST
12:30 a.m. EDT	18:13 GMT

Table 4.1 Examples of "Time"

4.2.1.2 Date

A date expression is used to express certain date on the time line. There are two kinds of date expressions: temporal words, and composite expressions. Some words or phrases

present a time concept, such as “past”, and “now”. These temporal words do not mean specific year, month, week or day, but refer to common temporal concepts on the time line. We record the words and their corresponding meanings in dictionaries, therefore the system can recognize them normally.

Temporal expressions of the kind “Composite” consist of certain kinds of components, such as year, month, day, weekday, direction (ago, later), etc. They can be divided into two types: precise expressions (ISO temporal expressions and NON-ISO temporal expressions) and calculated expressions. ISO format expressions are subset of the expressions defined by ISO 8601 standard, using a sequence of numbers to denote a date. Non-ISO expressions consist of numbers and temporal units and they are be combined according to certain patterns. For example, “1998-07-05” is an ISO expression, while “April 28, 1999” and “Monday, October 19” are Non-ISO basic expressions.

Sometimes people use an indirect temporal representation to denote a date, such as “the day before yesterday”, “after two years” and “in the year 1999”. These expressions are named calculated date expression and they all have three basic elements: a reference time, a direction and an offset. There are three kinds of time directions, going backward, going forward and standing. From the reference time, along with the direction and move the offset then we can get the value of a calculated expression. Typical expressions of the kind “Date” are listed as follows.

May 16, 1996	Thursday March 28
now	the 1960s
January 10	the day before yesterday
April 1968	'30s
January 24, 2000, Monday	11 Dec 1998

Table 4.2 Examples of “Date”

4.2.1.3 Duration

A duration expression indicates how long a period is. Most of duration expressions are the combinations of numbers and temporal units, some of them can’t be anchored on the

time line, such as example, “two months”, and “three years”. Other duration expressions can be anchored on the time line. Normally they consist of a reference time, a direction and the lasting time, such as “the next two years” and “the past two weeks”. In Chinese some duration expressions and some date expressions are all the combinations of numbers and time units, disambiguation should be considered according to the context. However, In English date expressions and duration expressions normally have different representations, either by different words or different patterns. Typical expressions of the kind “Duration” in English are listed as follows.

a week	20 minutes
72 hours	two weeks
50 years	the next two years
16 months	the past two weeks
three months	more than 40 years

Table 4.3 Examples of “Duration”

4.2.1.4 Pos_ Date

People always append a general description to denote a position in a larger scope, if they do not know the exact number at an inferior time level, such as “the spring of last year”, “morning, 9 May”, “next afternoon”. These expressions include date expressions and imprecise appendixes. The appendixes can be appended only once in the expression. In addition, the granular level of the appendixes should be given according to the minimal granularities of the temporal expressions. For example, “spring” follows “year” and “morning” follows “day”. According to different date unit levels, we can classify them into two groups: position of year and position of day. They are shown in Table 4.4.

morning	tomorrow morning
spring	yesterday afternoon
afternoon	Friday morning
this summer	next afternoon
night	Monday morning

Table 4.4 Examples of “Position of Date”

4.2.1.5 Frequency

These expressions consist of duration expressions and frequency indicators, such as “every”, “each” and “per”. They are used to measure the fixed time intervals between two repeating events, such as “every year”, “every two days” and “per week”.

4.2.1.6 Event-anchored

Event-anchored expressions are about the time of events, such as “when he was speaking”, “two days before the time he was born”. These expressions may consist of the description of events and time indicators, such as “when”. Event-anchored expressions can be anchored on the time line by reference times, the directions of offsets and the values of offset. The difference between event-anchored expressions and calculated expressions is that the reference time of event-anchor expressions are the times of events, not values of date expressions.

4.2.2 English Temporal Expression Extraction

The extraction task is to identify the extents of temporal expressions in the surface texts. Given a document, chunks of text which are temporal expressions should be marked out. A document with tagged temporal expressions is shown in Figure 4.2. A set of context free grammar rules is designed to describe the forms of all kinds of temporal expressions and a bottom-up chart parser is built to parse the temporal expressions. We have discussed types of temporal expressions in Section 4.2.1. Different grammar rules are designed for different types of expressions. To recognize basic components of temporal expressions, multiple dictionaries are built. Constraint rules are designed to distinguish temporal from non-temporal expressions according to the context and they will be applied before corresponding grammar rules are conducted. As the char parser keep each possible sequence of words, including nested and adjacent temporal expressions, combination rules are designed to combine the temporal expressions extracted by grammar and constraint rules. The final result

is the longest temporal expression. The dictionaries, grammar rules, constraint rules and combination rules are described in Section 4.2.2.1, Section 4.2.2.2, Section 4.2.2.3 and Section 4.2.2.4 respectively.

<DOC>
<DOCNO> CNN20001025.1400.0281 </DOCNO>
<DOCTYPE> NEWS STORY </DOCTYPE>
<DATE_TIME> **<TIMEX2>10/25/2000 14:04:41.42</TIMEX2>**
</DATE_TIME>
<BODY>
<TEXT>
Russian and Norwegian divers have succeeded in cutting a hole in the sunken submarine "kursk." it took the divers **<TIMEX2>five days</TIMEX2>** to cut through the 16-inch hull. So far three bodies have been found. The nuclear-powered sub sank on **<TIMEX2>August</TIMEX2>** killing 118 Russia sailors. Experts say the recovery may be long, risky and expensive. At least six more holes will have to be cut into the sub to allow access to each sealed-off compartment. Russian officials blame a collision with another sub and promised to have proof within **<TIMEX2>the next two months</TIMEX2>**.
</TEXT>
</BODY>
<END_TIME>
<TIMEX2>10/25/2000 14:05:14.02</TIMEX2>
</END_TIME>
</DOC>

Figure 4.2 Temporal expressions marked in an English document

4.2.2.1 Dictionaries

Dictionaries are designed to identify basic components in temporal expressions, such as month, weekday, direction and temporal unit. Each possible sequence of words in a sentence will be checked, if it match a word or phrase in certain dictionary, then the corresponding component is identified. For example, given the sentence “It took the divers five days to cut through the 16-inch hull”, the components “number” (five) and “temporal unit” (days) are recognized. Thirty four dictionaries are designed in our system. One dictionary “Month” is shown as follows. All possible words which represent meanings about “month” are recorded in this dictionary, such as “January” and “July”. It can be seen that there are multiple different representations for the same month. For example, “January”, “Jan.” and “JAN” all

refer to the first month of a year. The procedure to build dictionaries is rather time consuming.

Entry	Month		Entry	Month
January	1		July	7
Jan	1		Jul	7
Jan.	1		Jul.	7
JAN	1		JUL	7
JAN.	1		JUL.	7
…	…		…	…

Table 4.5 The dictionary “Month”

4.2.2.2 Grammar Rules

A set of grammar rules is designed for each kind of temporal expressions. In order to cover more temporal expressions, the rules are rather loose. It is unavoidable to extract pseudo temporal expressions along with true temporal expressions. This problem will be addressed in Section 4.2.2.3. The grammar rules to recognize a kind of “Date” expressions are shown in Table 4.6.

No.1. Exp -> Date
No.2. Date -> Composite
No.3. Composite -> Precise
No.4. Precise -> Non_ISO_Format
No.5. Non_ISO_Format -> Num_Base
No.6. Num_Base -> Numeral +

Table 4.6 Examples of English grammar rules

Given a temporal expression “2007”, first grammar rule No. 6 in Table 4.6 is applied and the text “2007” is recognize as the “Num_Base”, as “2007” is a sequence of numerals. Then grammar rule No. 5, No.4, No.3, No.2 and No.1 in Table 4.6 are applied sequentially. Finally, “2007” is identified as a temporal expression. One grammar rule is enough to identify this expression, e.g. “Exp -> Numeral +”, but there are kinds of temporal expressions and we have to organize them in a tree style scheme. The component “Date”, “Composite”,

“Precise”, “Non_ISO_Format” are kinds of temporal expressions and “Num_Base” is a kind of number.

4.2.2.3 Constraint Rules

Complexity and variety are pervasive in natural languages. Even when the domain is narrowed down to a particular field, like the temporal processing, the grammar rules still can’t guarantee to match only with those true temporal expressions. Pseudo expressions are those who match the grammar rules but are non-temporal in nature. Constraint rules are designed to distinguish temporal from non-temporal expressions according to the context. These constraint rules are developed manually.

The constraint rule is triggered when the right part of the corresponding grammar rule is matched. A grammar rule can be applied if and only if the associated constraint rule is satisfied. Examples of the constraint rules are given in Table 4.7. The following two examples then illustrate the constraint checking procedure step by step.

Grammar rule 5:	Non_ISO_Format -> Num_Base
Constraint rule 5:	IF the value of “Num_Base” is between 1800 and 2100, THEN the constituent “Num_Base” can be regard as Non_ISO_Format -- a kind of temporal expressions.

Table 4.7 Examples of English constraint rules

(1) The distance between the two cities is 1287 kilometers long.

Step 1. Recognize numbers.

[1/Numeral][2/Numeral][8/Numeral][7/Numeral]

Step 2. Apply the grammar rule No.6.

[1287/Num_Base]

Step 3. Check constraint rule No.5.

Fail and then terminate parsing.

(2) This news agency reported the event at 1996.

Step 1. Recognize numbers.

[1/Numeral][9/Numeral][9/Numeral] [6/Numeral]

Step 2. Apply the grammar rule No.6.
[1996/Num_Base]
Step 3. Check constraint rule No.5.
Pass.
Step 4. Apply grammar rule No.5.
[1996/Non_ISO_Format]
Step 5. Apply grammar rule No.4.
[1996/Precise]
Step 6. Apply grammar rule No.3.
[1996/Composite]
Step 7. Apply grammar rule No.2.
[1996/Date]
Step 8. Apply grammar rule No.1.
Recognize the temporal expression successfully.

In the first example the numeral sequence "1287" is not a temporal expression. However, the sequence "1996" can be recognized as a year, as in the second example. The constraint rule 5 is thus necessary for filtering out the pseudo expression "1287" in example (1).

4.2.2.4 Combination Rules

Since all possible word sequences in a sentence are examined, multiple nested, overlapped and adjacent temporal expressions will be recognized. Maybe they are parts of the complete temporal expressions. Combination rules are applied to integrate the temporal expressions extracted by grammar and constraint rules. The combination procedure is demonstrated by the following two examples.

(3) This news was published at April 28, 1999.
First recognized temporal expressions are [April], [April 28], [1999] and [April 28, 1999]. After the combination, the correct answer [April 28, 1999] will appear.

(4) The basketball game will start at February 10, Thursday.
First recognized expressions are [February], [February 10], [Thursday] and [February 10, Thursday]. The final result is [February 10, Thursday].

4.2.2.5 Temporal/Non-Temporal Disambiguation

Some word sequences are temporal expressions in given contexts, but not in other contexts. The context must be taken into account in order to facilitate the temporal/non-temporal disambiguation. Some constraint rules are designed for disambiguation purpose.

Three kinds of ambiguities have been founded. The first kind is caused by numbers, as in the example (5). In this case, the expression "15：10" contains temporal information, but it may be the score of a game in sport news. The second kind is caused by certain English words, such as "former". In example (6), it means "occurring earlier in time". But it means "in front of" in the context of "coming before in order".

(5) This train will arrive at the terminal at 15:10.
(6) Yeltsin, the former Russia president, and other officials wrote the sentences for human peace.

Ambiguities come forth when more than one interpretation is applied to the same phrase or word. To discriminate these expressions, heuristics are designed in the corresponding constraint rules.

4.2.3 English Temporal Expression Normalization

Attributes	Functions
VAL	Contains the value of a time or duration
MOD	Captures temporal modifications
SET	Designates frequency expressions
ANCHOR_VAL	Contains a normalized form of the reference time
ANCHOR_DIR	Capture the relative direction/orientation between VAL and ANCHOR_VAL
NON_SPECIFIC	Designates a generic, essentially non-referential expression

Table 4.8 Temporal attributes

The TERN 2004 evaluation is a public evaluation on extraction and normalization of temporal expressions. To evaluate our algorithms in a real task, we express temporal

information according to the guidelines of TERN 2004 evaluation. Any temporal expression is normalized as a possible combination of the six attributes in Table 4.8. Given the temporal expression “next afternoon”, the normalized result should be VAL =“2000-10-07TAF” and the other attributes are left blank. Here “2000-10-07” is the date of “tomorrow”. As we have discussed in previous sections, different kinds of temporal expressions have different normalization results. They are detailed in the following sections.

4.2.3.1 Temporal Object Involved

Rule based normalization is based on the parsing results which have been obtained during extraction. Normalization can be regarded as a mapping procedure that interprets the temporal expressions with the temporal attributes. It attempts to “understand” the “temporal meanings”. First of all, we introduce temporal objects and measurement which are necessary in the normalization procedure.

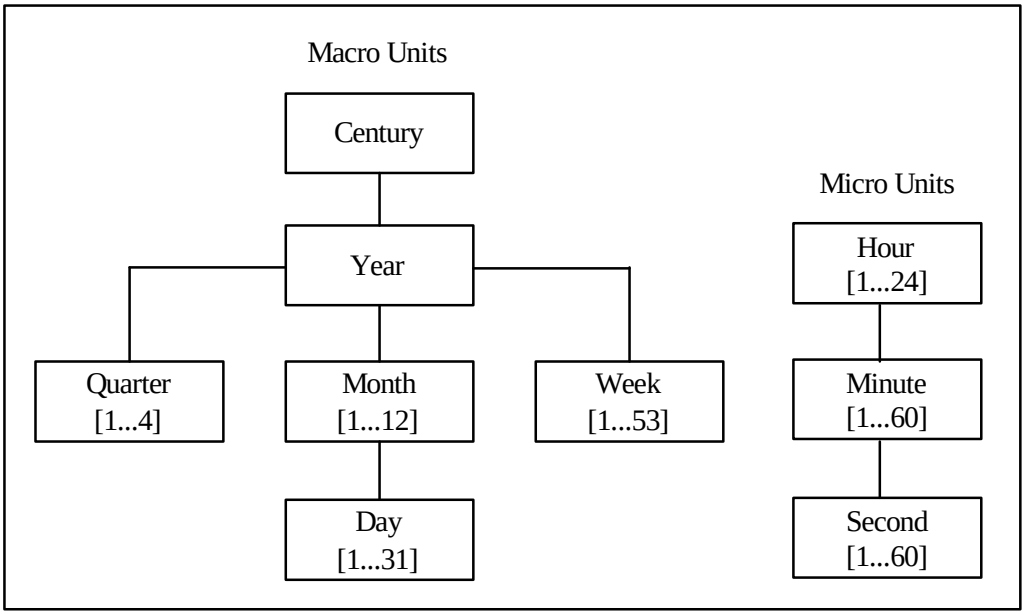

Figure 4.3 The scheme of temporal units

In the field of time, the basic objects are time and durations. Time is represented by points or intervals on the time line. Given the origin and a measurement, it can be evaluated with a natural number. If there is no extra specification, the calendar is the Gregorian. Duration is the distance between two times. One can anchor duration by its starting time and

end time, or by one of them and the length of the duration. However, when duration refers to length, it can't be anchored on the time line. The temporal relations between two objects, such as "before", "same", "include" and "after", are the intrinsic concepts behind surface texts. They should be discovered in the normalization as well.

To represent lengths on the time line, the measurement is needed. The temporal units are either macro or micro units, which are shown in Figure 4.3. To represent a time, the scope of the numbers which can be combined with the temporal units is limited. "Century" and "year" are two special time units, because only these two time units can anchor a time concept on the time line. Without clues in contexts, other time units can't anchor a time concept on the time line individually.

4.2.3.2 Rule-Based Normalization

The chart parser keeps all applied grammar rules and recognized intermediate constituents during extraction. Temporal semantics of the extracted expressions can be obtained by examining these rules. Some basic objects, such as "number", "unit", "time" and "duration", are employed to store temporal information. The procedure of normalization is to create or update these objects. This procedure can be regarded as mapping temporal expressions to the six attributes. Different mapping procedures are applied to different kinds of temporal expressions. A general description is shown in Figure 4.4.

"MOD" attribute is set to "YES" if the expressions are modified by "about" and "before", etc. Any kind of temporal expressions may have this attribute. "Frequency" expressions can be explained as set of times, such as "each year" and "per week", or set of durations, such as "every two years". For the temporal expression "every two years", the attributes should be assigned as "VAL = 2Y, SET = YES". "ANCHOR_VAL" and "ANCHOR_DIR" refer to reference times. We pick up the publication time of a news article as the default reference time.

Event-anchored temporal expressions are relevant with some specific events, such as "When visiting U.N. headquarters in New York" and "when talking with U.S. congressmen".

The human annotating results about event-anchored expressions are irregular. Some expressions are annotated as blank, while others are annotated as the days on which the events occurred. It is hard to build and tune the rule set according to the golden answers. Therefore, in our system event-anchored expressions are not normalized. Since it is hard to tell whether a temporal expression is “specific” or not even for human, the attribute “NON_SPECIFIC” is ignored. Actually, only very few expressions have this attribute. Six selected examples are provided in Table 4.9.

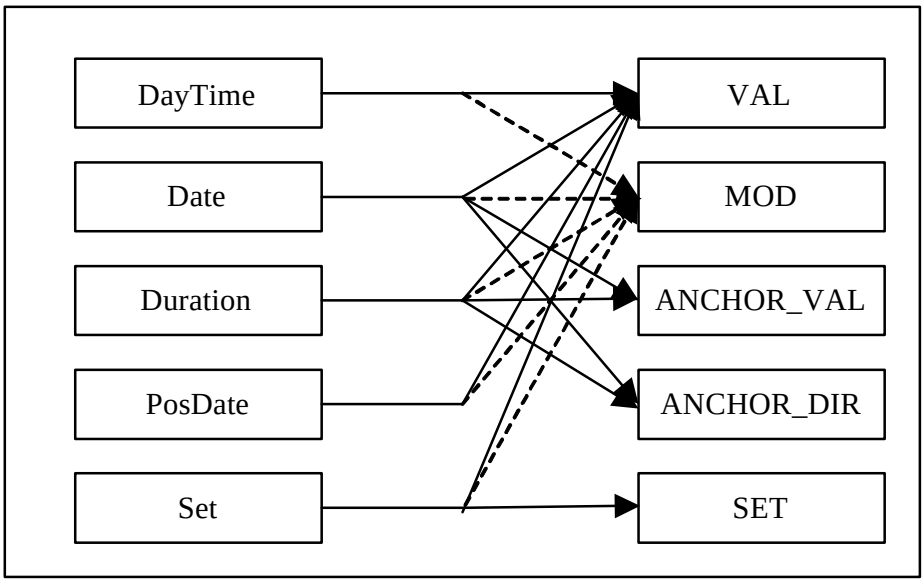

Figure 4.4 Mapping temporal expressions to attributes

Expressions	Attributes
20: 20 p.m.	VAL="1999-04-26 T20:20"
the next year	VAL="1999"
every two days	VAL="P2D" SET="YES"
next afternoon	VAL="2000-10-07TAF"
about ten days	VAL="P10D" MOD=”APPROX”
now	VAL="PRESENT_REF" ANCHOR_VAL="2000-10-05" ANCHOR_DIR="AS_OF"

Table 4.9 Examples of English temporal expressions normalized

4.3 Chinese Temporal Expression Processing

4.3.1 Chinese Temporal Expression Classification

According to our observation on Chinese texts and the guideline of TERN 2004, Chinese temporal expressions can be classified into different classes. They are shown in the Figure 4.5.

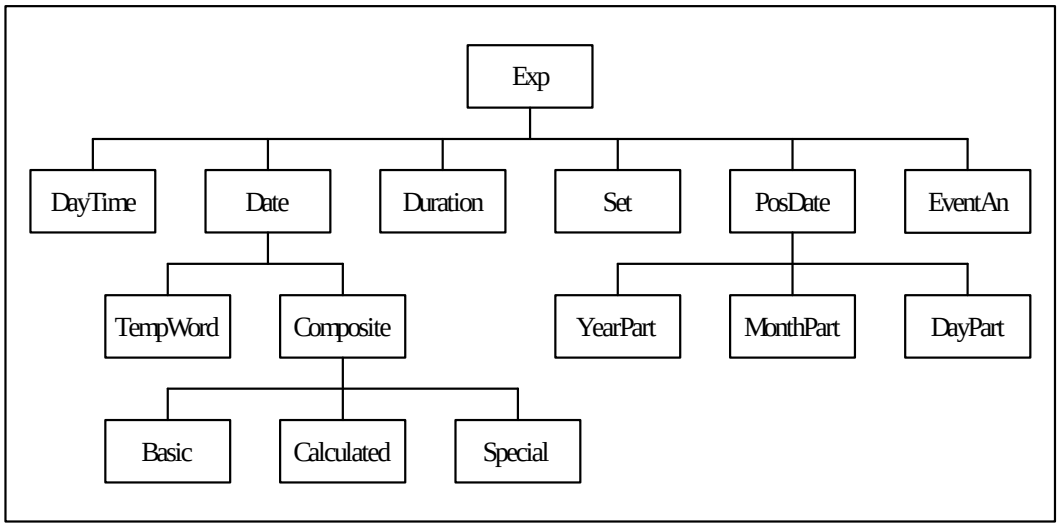

Figure 4.5 The classification of Chinese temporal expressions

In Chinese, if people do not know the exact representation at an inferior temporal level, they may append an imprecise description to denote a position in a larger scope, such as “去年春天/the spring of last year”. These temporal expressions are named “PosDate”. These expressions consist of date expressions and imprecise appendix.

“TempWord” expressions are some Chinese words which contained temporal meanings, such as “春节/the lunar new year”, “目前/now”. “Composite” expressions include basic temporal expressions, calculated expressions and special expressions, such as “1999 年 4 月 28 日/April 28, 1999”, and “两年后/after two years” and “1999 财年/ the fiscal year 1999”. “Set” expressions denote a set of time and most of them are about frequency, such as “每年/every year” and “每两天/every two days”. “EventAn” expressions are relevant to the times

of events, such as “当他演讲时/when he was speaking”. “EventAn” expressions can be anchored on the time line only after the times of the events are resolved.

4.3.2 Chinese Temporal Expression Extraction

The task of extraction is to identify the extents of temporal expressions in the surface text. A set of context free grammar rules is designed to describe the basic form of all kinds of Chinese temporal expressions and a bottom-up chart parser is employed to parse temporal expressions. Word segmentation is a preliminary step in many Chinese NLP applications. However, the performance of word segmentation is not perfect and it may introduce some errors. In our system, each possible combination of Chinese characters in a sentence will be looked up, and then all of the constituents are fed into the char parser. If the dictionaries are comprehensive enough, then all the possible combinations of characters can be gotten. Ambiguities and overlaps between multiple temporal expressions are left to constraint rules and combination rules.

4.3.2.1 Grammar Rules

A set of grammar rules is designed for each type of Chinese temporal expressions. In order to catch more temporal expressions, the grammar rules are given loosely. Some pseudo temporal expressions may be introduced and this problem is addressed in the next section. Given the grammar rules of table 4.10, “15 时 24 分/15:24” and “15 时 24 分 39 秒 /15:24:39” can be recognized. In these examples, “时/o’clock”, “分/minute”, “秒/second” are constituents of the type “Time_Unit”.

No.1. Exp -> Time_Of_Day
No.2. Time_Of_Day -> Time_Base
No.3. Time_Base -> Time_Temp +
No.4. Time_Temp -> Integer Time_Unit
No.5. Integer -> Digit +

Table 4.10 Examples of Chinese grammar rules

4.3.2.2 Constraint Rules

There are many complex and variable phenomena in natural language. Even if the domain is narrowed down to the temporal field, grammar rules are not enough to extract exact temporal expressions. There are some pseudo expressions which satisfy grammar rules, so constraint rules are designed to filter these expressions according to the context. These constraint rules are developed by analyzing the data set.

A constraint rule will be triggered after the right part of the corresponding grammar rule is satisfied. If the constraint rule is satisfied, then the grammar rule can be applied; otherwise, it cannot be applied. Examples of constraint rules are shown in Table 4.11 and the following two examples show the constraint checking procedure step by step.

Grammar rule 3:	Time_Base -> Time_Temp +
Constraint rule 3:	IF There is only one constituent of the type “Time_Temp”, THEN the constituent “Time_Unit” which is contained in “Time_Temp” should not be “分/minute”.
Grammar rule 4:	Time_Temp -> Integer Time_Unit
Constraint rule 4:	The constituent “Integer” can not end up with “个/ (a quanti fier)”.

Table 4.11 Examples of Chinese constraint rules

(5) 这家新闻机构十分迅速地报道了这次事件。

(This news agency reported the event very quickly.)

Step 1. Look up dictionary.

[十/Digit] [分/Time_Unit]

Step 2. Apply the grammar rule No.5.

[十 /Integer] [分/Time_Unit]

Step 3. Check constraint rule No.4.

Pass.

Step 4. Apply grammar rule No.4.

[十分/Time_Temp]

Step 5. Check constraint rule No.3.

Fail and then terminate parsing.

(6) 晚上 7 时 30 分西弗吉尼亚州和俄亥俄州投票结束。

(The ballot ended at 7:30 p.m. in Western Virginia and Ohio.)

Step 1. Look up dictionary.

[7/Digit] [时/Time_Unit][3/Digit] [0/Digit][分/Time_Unit]

Step 2. Apply the grammar rule No.5.

[7/Integer] [时/Time_Unit] [30/Integer][分/Time_Unit]

Step 3. Check constraint rule No.4.

Pass.

Step 4. Apply grammar rule No.4.

[7 时/Time_Temp] [30 分/Time_Temp]

Step 5. Check constraint rule No.3.

Pass.

Step 6. Apply grammar rule No.3.

[7 时 30 分/Time_Base]

Step 7. Apply grammar rule No.2.

[7 时 30 分/Time_Of_Day]

Step 8. Apply grammar rule No.1.

[7 时 30 分/Exp]

Step 9. Recognize the temporal expression successfully.

In the first example “十分/very” is an adverb and has no temporal meaning. However the character “十/ten” and “分/minute” can be looked up and satisfy the grammar rule. Constraint rules are necessary to filter this pseudo expression.

4.3.2.3 Combination Rules

Because each possible substring in a sentence is tried, multiple nested, overlap and adjacent temporal expressions may exist in the sentence. However, some of these expressions are just parts of the optimal answers. So combination is necessary to get the integrated temporal expression. After applying grammar rules, if any two temporal expressions are nested, overlapped or adjacent, our system will combine them and keep the final result. This procedure is shown by the following examples.

(7) 这次列车将于次日早上到达南昌。

(This train will arrive at Nan Chang next morning.)

First recognized temporal expressions are [次日/tomorrow] and [早上/morning].

After the combination, the correct answer [次日早上/next morning] will appear.

(8) 晚上 8 时篮球比赛开始。

(The basketball game starts at 8:00 p.m.)

First recognized expressions are [晚上/night], [8 时/8:00], [晚上 8 时/8:00 p.m.].

The final result is [晚上 8 时/ 8:00 p.m.].

4.3.2.4 Temporal/Non-Temporal Disambiguation

Some strings of characters are temporal expressions in given contexts, but in other contexts they are not. The context should be browsed to extract the true temporal expressions. Some constraint rules are designed to check the context and fulfill disambiguation. Three kinds of ambiguities are founded. The first kind is the ambiguities caused by numbers, such as example (9). In this case, the expression "15：10" contains temporal information, but in sports news messages it may be a score of a game. The second kind is the ambiguities caused by the combination of numbers and time units, such as "10 号". In example (10), the expression "10 号" just refers to a football team member. However, in many news messages it is a date. The third kind is the ambiguities caused by Chinese words, such as "前". In example (11), the expression means "former" and its explanations in other contexts may be "in front of".

(9) 本次列车将于 15：10 到达终点站。

(10) 然而 6 分钟后，10 号宿茂臻冲顶即将比分扳平。

(11) 俄罗斯前总统叶利钦等政府首脑为人类和平欣然提笔。

There are multiple explanations for the same one phrase or word, so ambiguities may be caused. To discriminate these expressions, heuristics for disambiguation are embedded in corresponding constraint rules.

4.3.3 Chinese Temporal Expression Normalization

The goal of normalization is to represent the temporal information contained in temporal expressions, according to certain standard. Normalization is based on the mapping procedure, in which temporal expressions are explained and represented by values of temporal attributes. In this procedure, the objects number, unit, time and duration are employed to store and represent temporal information. The temporal attributes involved in Chinese temporal expression normalization are same with those in English temporal expression normalization.

After the procedure of extraction, the chart parser keeps all the applied grammar rules and recognized intermediate constituents. Semantic meanings of temporal expressions can be achieved by the explanation of these grammar rules. In this procedure, some basic objects, such as “number”, “unit”, “time” and “duration”, can be employed to store and convey temporal information. Applying grammar rules means creations or updates of basic temporal objects. Based on our temporal framework, we explain how to normalize the temporal expression extracted, i.e. mapping the expressions to the values of six attributes. The mapping procedure is different for different kinds of temporal expressions. A general description about the mapping procedure is shown in Figure 4.4.

According to the Chinese classification scheme in Section 4.3.1, all temporal expressions can be mapped to the six attributes. The mapping procedures are complicated and selected examples are shown in table 4.12. It is difficult to tell whether a temporal expression is “specific” or not, and few expressions are set a value for this attribute, we do not map expression to the attribute “NON_SPECIFIC”. “MOD” attribute of temporal expressions may be set as “YES” if there are some modifying descriptions about the expressions, such as “将近/about”, “早于/before” and so on. So any kind of temporal expressions may be mapped on this attribute. “Set” expressions can be explained as set of times, such as “每年/each year”, or set of durations, such as “每两年/every two years”, so the attributes “VAL” and “SET” will be filled. “ANCHOR_VAL” and “ANCHOR_DIR” refer to reference times and

we adopt the publishing times of news articles as the default reference times. Event expressions are relevant with specific events and it is hard to represent the exact meaning of them. In our system event expressions are not normalized.

Expressions	Attributes
目前/now	VAL="PRESENT_REF" ANCHOR_VAL="2000-10-05" ANCHOR_DIR="AS_OF"
晚上 8 时 20 分/20: 20 p.m.	VAL="1999-04-26 T20:20"
后两年/the next two years	VAL="P2Y" ANCHOR_VAL="2000" ANCHOR_DIR="STARTING"
每两天/every two days	VAL="P2D" SET="YES"
明天下午/ next afternoon	VAL="2000-10-07TAF"

Table 4.12 Examples of normalized Chinese temporal expressions

4.3.3.1 Time/Duration Disambiguation

Sometimes people omit a part of a full temporal expression for convenience in Chinese texts. For example, “4 月/April” and “97 年/ ’97” are used to instead “2000 年 4 月/April, 2000” and “1997 年/the year 1997”. However, “4 月/four months” and “97 年/97 years” are also legal temporal expressions by themselves. These temporal expressions are combinations of numbers and temporal units. The first kind of explanations means that these expressions are times and the second kind of explanations means that they are durations. To fill correct values in temporal attributes for these temporal expressions, disambiguation is necessary. Heuristic rules are employed for disambiguation. Examples of disambiguation rules are shown in table 4.13.

IF a 3-digit or four-digit number is combined with the unit “年/year”, THEN this expression is time; IF a 2-digit number is combined with the unit “年/year” and the number is bigger that 70, THEN this expression is time. IF a 1-digit number is combined with the unit “年/year”, THEN this expression is duration.

Table 4.13 Examples of Chinese disambiguation rules

4.4 Experiment and Evaluation

To evaluate the approaches for English temporal expression extraction and normalization, we choose a manually annotated corpus, which consists of 511 English documents. They come from the documents used for ACE 2002, 2003 and 2004 evaluations. 5324 temporal expressions are annotated in this data set. After implementing our temporal expression extraction and normalization system, we evaluated it on this corpus with a scorer program provide by TERN 2004. To evaluate the approaches for Chinese temporal expression extraction and normalization, we choose a manually annotated corpus, which consists of 457 Chinese news documents. The data collection contains 285,746 characters/142,872 words and 4,290 manually annotated Chinese temporal expressions.

4.4.1 Experiments on English Temporal Expression Extraction

In this section, we report the evaluation results on rule based temporal expression extraction. Our temporal processing system embeds the constraints to restrict the grammar rules, and combines the nested, overlapped and adjacent temporal expressions. After the system is evaluated by the scorer provided by TERN 2004, a performance report is generated. Among total 5324 expressions, our system identify 3652 correctly and 554 incorrectly. In addition, the system misses 1118 expressions and output 242 spurious expressions. Therefore the precision, recall and F-measure [Salton and McGill 1983] are 0.821, 0.686 and 0.747 respectively.

For reference, we compared the evaluation results of our system with those of other systems in TERN 2004 (Table 4.14). Please note that the data set used for TERN 2004 final evaluation is similar with the data set employed in our experiments, but not same.

System	Precision	Recall	F-measure
A	0.872	0.827	0.849
B	0.885	0.798	0.839
C	0.901	0.681	0.776
D	0.687	0.567	0.621
E	0.830	0.451	0.584
F	0.290	0.237	0.261
Our system	0.821	0.686	0.747

Table 4.14 Evaluation results on English temporal expression extraction

4.4.2 Experiments on English Temporal Expression Normalization

We conduct experiments on normalization on the same data set and evaluate the system performance with the scorer. Table 4.15 presents the performance of our system. From it we can see recall scores are lower than the precision scores. The possible reason is that the coverage of our grammar rule set is limited. We plan to tune the system in future to improve the performance further, especially on the aspect of recall. We then compare our normalization results with those of other systems evaluated in TERN 2004 according to different temporal attributes, such as VAL, MOD, SET, ANCHOR_DIR and ANCHOR_VAL. The performance is shown in Table 4.16, Table 4.17, Table 4.18, Table 4.19, and Table 4.20 respectively. Our performance is medium among theirs in general.

	Total	Corr	Inco	Miss	Spur	Prec	Rec	F
VAL	4192	3372	714	106	7	0.824	0.804	0.814
MOD	164	72	6	86	42	0.600	0.439	0.507
SET	86	47	0	39	2	0.959	0.547	0.686
ANCHOR_DIR	749	434	3	312	3	0.986	0.579	0.730
ANCHOR_VAL	749	302	133	314	3	0.689	0.403	0.509

Table 4.15 Evaluation results on English temporal expression normalization

System	Precision	Recall	F-measure
A	0.866	0.837	0.851
B	0.875	0.870	0.872
C	0.843	0.847	0.845
D	0.686	0.709	0.698
E	0.798	0.640	0.710
F	0.671	0.674	0.673
Our System	0.824	0.804	0.814

Table 4.16 Evaluation results according the temporal attribute VAL

System	Precision	Recall	F-measure
A	0.600	0.058	0.105
B	0.837	0.720	0.774
C	0.840	0.553	0.667
D	0.444	0.111	0.178
E	0.000	0.000	0.000
F	0.061	0.211	0.094
Our System	0.600	0.439	0.507

Table 4.17 Evaluation results according the temporal attribute MOD

System	Precision	Recall	F-measure
A	0.974	0.776	0.864
B	0.880	0.564	0.688
C	0.000	0.000	0.000
D	0.882	0.455	0.600
E	0.000	0.000	0.000
F	1.000	0.250	0.400
Our System	0.959	0.547	0.696

Table 4.18 Evaluation results according the temporal attribute SET

System	Precision	Recall	F-measure
A	0.621	0.612	0.617
B	0.833	0.698	0.760
C	0.986	0.371	0.539
D	0.818	0.566	0.669
E	0.851	0.312	0.457
F	0.000	0.000	0.000
Our System	0.986	0.579	0.730

Table 4.19 Evaluation results according the temporal attribute ANCHOR_DIR

System	Precision	Recall	F-measure
A	0.657	0.748	0.700
B	0.683	0.775	0.726
C	0.986	0.371	0.539
D	0.703	0.487	0.575
E	0.041	0.015	0.022
F	0.000	0.000	0.000
Our System	0.689	0.403	0.509

Table 4.20 Evaluation results according the temporal attribute ANCHOR_VAL

4.4.3 Experiments on Chinese Temporal Expression Extraction and Normalization

In this section we report the results about evaluating our Chinese temporal expression extraction and normalization approaches on a manually annotated Chinese corpus. We evaluate the boundaries of expressions and the values of the six temporal attributes.

Experiment No.	Conditions
1	No constraints, combination of nested expressions
2	No constraints, combination of nested, overlapped and adjacent expressions
3	Constraints, combination of nested expressions
4	Constraints, combination of nested, overlapped and adjacent expressions

Table 4.21. Experimental configuration for Chinese temporal expression extraction and normalization

Constraints have been embedded in the system to restrict grammar rules. In addition, nested, overlapped and adjacent temporal expressions have been combined. In Chinese, many temporal expressions contain nested temporal expressions. If these nested components are not combined into optimal answers, there will be so many extra mismatched expressions. Therefore the combination of nested temporal expressions is necessary. In the experiments, we try to evaluate two factors: constraint rules, and the combination of overlapped and adjacent temporal expressions. Four experiments are set up, which are described in Table 4.21. Given these configurations, the results of the experiments are shown in Table 4.22.

Attributes		No. 1	No. 2	No. 3	No. 4
TEXT	P	0.717	0.758	0.810	0.856
	R	0.838	0.850	0.830	0.843
	F	0.773	0.801	0.820	0.849
VAL	P	0.730	0.750	0.787	0.807
	R	0.693	0.681	0.742	0.732
	F	0.711	0.714	0.764	0.768
MOD	P	0.563	0.565	0.629	0.626
	R	0.586	0.550	0.616	0.574
	F	0.574	0.557	0.622	0.599
SET	P	0.698	0.662	0.879	0.867
	R	0.606	0.589	0.611	0.598
	F	0.649	0.624	0.720	0.707
ANCHOR_VAL	P	0.680	0.750	0.681	0.687
	R	0.658	0.681	0.662	0.652
	F	0.669	0.714	0.672	0.669
ANCHOR_DIR	P	0.724	0.727	0.733	0.737
	R	0.682	0.669	0.694	0.682
	F	0.702	0.697	0.713	0.708

Table 4.22. Evaluation results on Chinese temporal expression extraction and normalization

Several related works are designed to extract and normalize temporal expressions, but they are about English, Spanish, French, and Korea. We take part in TERN 2004 evaluation on Chinese temporal expression extraction and achieve the highest performance in this task.

There is no public result on Chinese temporal expression normalization, for reference we compare our normalization result of Experiment No.4 with the English normalization result in TERN 2004. Our performance is medium among their results.

Table 4.22 compares the Precision, Recall and F-measure for different attributes in different experiments. “TEXT” means the performance of exact boundaries of temporal expressions and other attributes are explained in Section 4.2.3. For attributes “TEXT” and “VAL”, we achieve the highest performance in Experiment No.4. The F-scores are 0.849 and 0.768, respectively. For other attributes, we also achieved nearly highest performance in Experiment No.4. From the trend of performance on these two attributes, it can be seen that the constraints and the combination procedure help to improve the performance on temporal expression extraction and normalization, especially on “TEXT” and “VAL”. At the same time, the combination procedure is not significant to other attributes. Based on the assumption that two adjacent or overlapped temporal expressions refer to the same temporal concept, we combined them. However, the combination procedure does not help to explain the meanings of temporal expressions.

After the evaluation we collect the errors of Experiment No.4 and try to find the reasons. Wrong attribute values include missed, incorrect and spurious cases. The reason for errors on the attributes “ANCHOR_VAL” and “ANCHOR_DIR” is that the system did not give correct reference times. Table 4.23 gives the error distributions according to different attributes. From this table, it can be seen that temporal Chinese words and events are difficult to extract and normalize.

Attributes	Reasons	Number	Percentage
TEXT	Boundaries of temporal Chinese words	366	37.4%
	Boundaries of events	193	19.7%
	Grammar rules	161	16.4%
	Boundaries of temporal noun phrase	89	9.1%
	Combination procedure	76	7.8%
	Annotation inconsistence	75	7.7%
	Temporal/non-temporal ambiguities	19	1.9%
VAL	Explained semantics	299	27.6%
	Explanation of temporal Chinese word	180	16.6%
	Errors introduced by extraction	177	16.3%
	Specification/generalization characteristic	148	13.7%
	Wrong reference times	122	11.3%
	Annotation inconsistence	80	7.4%
	Point/duration ambiguities	63	5.8%
	Explanation of events or noun phrase	14	1.3%
MOD	Errors introduced by extraction	44	33.3%
	Annotation inconsistence	35	26.5%
	Explanation of temporal Chinese word	27	20.5%
	Explained semantics	23	17.4%
	Ambiguities	3	2.1%
SET	Explained semantics	35	81.4%
	Errors introduced by extraction	3	7.0%
	Annotation inconsistence	5	11.6%

Table 4.23 The distribution of errors in Chinese temporal expression extraction and normalization

4.5 Chapter Summary

We investigate the rule based approach to extract and normalize comprehensive temporal expressions from English and Chinese texts. To cope with various temporal expressions, our rule based approach employs grammar and constraint rules to retrieve genuine expressions and resolve ambiguities. The two tasks are based on a powerful chart parsing and constraint checking scheme. In our experiments, the English rule-based approaches for exaction and

normalization are evaluated on a manually tagged corpus. We achieve promising results, i.e., F-measure 74.7% on extent and F-measure 81.4% on value. For reference the performance of this system is compared with that of other systems in TERN 2004 on similar data sets. In general our performance is medium among theirs.

We also investigate Chinese temporal expression extraction and normalization approaches. To cope with kinds of temporal expressions, constraint rules are employed to retrieve genuine expressions and resolve ambiguities. We have evaluated the extraction and normalization approaches on a manually annotated corpus and achieved promising results, i.e. F-measure 85.6% on extent and F-measure 76.8% on value. We take part in TERN-2004 Chinese temporal expression extraction with these approaches and achieve the highest performance in that track. From the experiments we find that constraints are significant to the task extraction and normalization. At the same time, combination has positive influence on the task extraction. Analysis shows that temporal Chinese words and events are more difficult to extract and normalize.

Chapter 5

Temporal-Oriented Event-Based Summarization

5.1 Chapter Overview

Summarization is a useful technique to help users identify important contents with bearable time cost. However, it is difficult to design a general summarization system which is applicable to all types of documents because of the limitation of NLP techniques nowadays. Therefore we focus on improve the performance of the summarization on certain type of text, such as, news reports with topics shifting over time. The work in Chapter 4 is the preparation for temporal-oriented summarization.

The key problems of summarization are how to represent documents and how to identify important contents. We consider that event is a natural unit to represent documents, especially for news reports. In general, an event can be described as “who did what to whom when and where”. Same as the definition presented in Chapter 3, event is defined as event terms together with the associated event elements (named entities) at sentence level. Event terms represent the actions themselves, including verbs and action nouns. Event elements denote event arguments, such as person names, organization names, locations and times. Given the sentence “Yasser Arafat on Tuesday accused the United States of threatening to kill PLO officials”, “accused”, “threatening” and “kill” are identified as an event term, while “Yasser Arafat”, “United States”, “PLO” and “Tuesday” are event elements. Encouraging

results of our event-based summarization approaches have been reported in Chapter 3. At that time, we don't consider the role of time in summarization.

We focus on news documents in this study, as they are formal and there are a lot of public data resources. Based on our observation, we find that topics may shift over time in news documents. Among 30 clusters of the DUC 2001 data set, about 10 clusters consist of descriptions on certain event happenings at different times, which are clearly declared in the original texts. For example, the theme of cluster d41 is "fires in USA". The fires in 1926, 1977, 1985, 1987 and 1990 are reported. We also analyse 102 queries of DUC 2005 and find 35 of them are temporal relevant. It can be seen that a substantial percentage of text involving in summarization is about time. This observation motivates us to investigate [Li et al. 2006a; Wu et al. 2007a] whether taking temporal distribution of events into account can improve the quality of summaries for these clusters in the context of event-based summarization.

After anchoring events on the time line, their importance can be evaluated from local and global points of views. Base on the observation, the important events are those occurring frequently in a certain period. Two statistical measures, i.e., *tf***idf* and χ^2, are explored. Either of them can be used to evaluate the importance of event terms and event elements based on their distribution on the time line. The weight of each sentence is calculated as the sum of the weights of all event terms and event elements contained in it. After evaluating importance of sentences, two kinds of sentence selection strategies are investigated, i.e. sequential and round robin selection. In our experiment, the combination of *tf***idf* and sequential sentence selection by sentence weight performs best. Compared with event-based summarization without considering temporal distribution, it improves ROUGE-1 by 18.8% on the two selected document clusters. In further evaluation on ten clusters of documents, this approach also achieves significant improvement when evaluated by human.

The remainder of the chapter is organized as follows. Section 5.2 introduces event representation on the time line. Section 5.3 then describes event weighting and sentence

selection strategies. Section 5.4 describes experiments and discusses the results. Section 5.5 analyses the reason of improvement from the introduction of temporal information and testifies the role of clustering strategy. Finally, Section 5.6 summarizes this chapter.

5.2 Event Representation on Time Line

Each sentence can be represented by a collection of instances of event terms and event elements. Event terms are verbs, and action nouns describing various types of actions, such as “election” and “extension”. Action nouns are extracted from WordNet hyponyms of “event” and “action”. Event elements are person names, organization names, locations and times. A public available tool, GATE [Cunningham et al. 2002], is used to identify the verbs functioning as event terms and the named entities functioning as event elements.

We have investigated documents representation by event instance and event concept in Chapter 3. In this chapter, the distribution of events on the time line is considered. Therefore the representation by event concept is not suitable, as it collapse different occurrences which refer to same concept into one single concept. We also investigate the influence of context on importance of event terms in Chapter 3, such as the number and types of neighboring named entities. It is found that the influence can be neglected. Therefore event terms and event elements are regarded as of same importance, i.e., we do not discriminate event elements and event terms when computing importance of them. Each event term/element will be represented on the time line and its importance will be evaluated later.

To represent event term/element on the time line, first the occurring time of events should be identified. As the purpose of this research is to investigate the role of time under the context of event-based summarization, we assign times to corresponding events manually first. TIMEX2 guidelines provide comprehensive descriptions about kinds of temporal expressions and their attributes. We tag event time according to it. Section 5.2.1 presents the procedure of time assignment in detail. Then each event instance, including event term and associated event elements, will receive their temporal values and can be anchored on the

time line. Considering time granularities mentioned with most news events are “day” or a coarser unit, we use “day” as the temporal unit to measure the time line (See Figure 5.1).

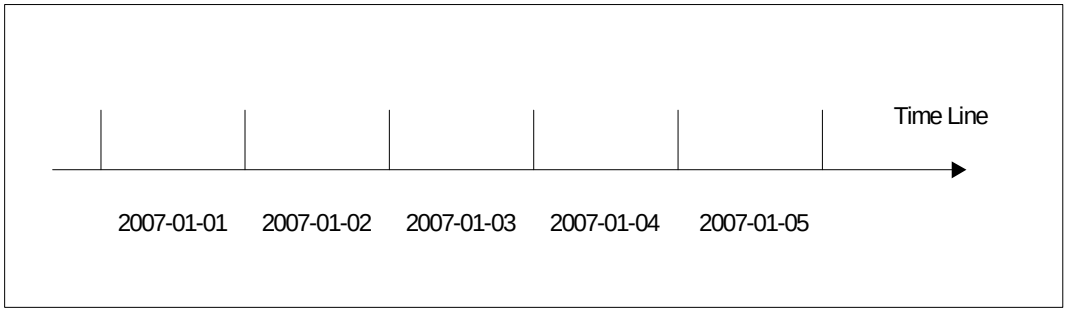

Figure 5.1 The time line measured by “day”

To evaluate importance of events we need to represent them with suitable forms. Events can be either instantaneous or durative. Given the sentence “The widespread drought of 1988 was replaced by spotty rain and local areas of dry weather”, this event is instantaneous. However, in the sentence “Californian fire has charred more than 1.3 million acres of forest and range land since January”, the event is durative. The difference between an instantaneous event and a durative event is how long the events last.

As Allen [1981] suggests, time interval can be used to represent both time point and duration, when points are regarded as intervals with “meeting places”. Thus on the time line, any instantaneous event can be represented by an interval with the same boundaries, which is denoted as a dot in this study. A durative event is then represented as an interval with two boundaries or just one boundary if the other one is unknown. According to the number of known boundaries, events can be classified into three groups, zero, one and two known boundaries. Section 5.2.2 and Section 5.2.3 describe the representations of these kinds of events in detail.

5.2.1 Event Time Assignment

After the identification of events in documents, we assign event times to them manually. Manual annotation is time-consuming, but high precision of temporal values of events can be guaranteed. To avoid errors from a temporal information processing system, which

automatically assigns temporal values to events, we decide adopt manual annotation first. Then we can focus on the role of time in event-based summarization. There are three steps to annotate a document. The first one is to identify the document time, at which the document is published. The second one is to judge boundaries of clauses, as different clauses may be associated different times. The third one is to normalize the time of the clause and assign temporal values to events in the clause.

<DOC>
<DOCNO> AP890801-0025 </DOCNO>
<FILEID>AP-NR-08-01-89 0300EDT</FILEID>
<TEXT>
<TIMEX2 val="PRESENT_REF" mod="" anchor_dir="AS_OF" anchor_val="1989-08-01">This week's flare-up of Western wildfires can't hold a candle to the damage wrought by last year's record-breaking fire season, but officials say a dry August could change everything.</TIMEX2>
<TIMEX2 val="P7M" mod="" anchor_dir="AFTER" anchor_val="1989-01">Fire has charred more than 1.3 million acres of forest and range land since January in the contiguous United States, compared to 2.1 million acres by this time last year, fire officials said Monday.</TIMEX2>
<TIMEX2 val="PAST_REF" mod="" anchor_dir="ON_OR_BEFORE" anchor_val="1989-08-01">``Right now, the fire season is just starting to gear up," said Sandi Sacher, spokeswoman at the federal government's wildfire command post in Boise, Idaho.</TIMEX2>
…
<TIMEX2 val="1988" mod="END" anchor_dir="" anchor_val="">By year's end, 6 million acres had burned in the West and Alaska, making 1988 the worst fire season in 30 years, and, in terms of firefighting resources committed, the most expensive in U.S. history, Sacher said.</TIMEX2>
<TIMEX2 val="PAST_REF" mod="" anchor_dir="ON_OR_BEFORE" anchor_val="1989-08-01">The widespread drought of 1988 has been replaced by spotty rain and local areas of dry weather, Sacher said.</TIMEX2>
….
</TEXT>
</DOC>

Figure 5.2 A document marked up with time

To keep the annotation consistent, a graduate is required to tag all the documents used in our experiment. The following standard is also set up for the assignment procedure. A tagged document is shown in Figure 5.2.

1. Event time is identified based on clause level. In one sentence, different event times may exist. In one clause, multiple events may exists, but they are assigned same time.

2. Temporal attributes and values are given according to TERN 2004 guidelines. Besides durations, time point or a period of time anchored on the time line may be assigned to events, such as “since January”, “Monday” and “nowadays”.
3. The default reference time is the document time. This temporal value can be extracted from the header tags of the document, such as <FILEID>. In the example document, the document time is “1989-08-01”.
4. If there is exact event time in a clause, it will be used for it. Otherwise, we judge the event time according to tense of the clause. If it is past, present or future tense, we assign temporal value of “past”, “now” and “future” to the clause respectively. The temporal value of “now” is shown as “<TIMEX2 val="PRESENT_REF" mod="" anchor_dir="AS_OF" anchor_val="1989-08-01">”.

5.2.2 Representation of Event with Two Boundaries

Events with two boundaries may be instantaneous or durative. For an instantaneous event or a durative event which last less than one day, as the unit of time line is “day”, it can be anchored in one day. To formulate the events occurred on a particular day (this information is clearly given in a sentence), we simply approximately represent these events by a dot on that day, and let the weight of each dot to be 1 (Figure 5.3, the upper part). It means we do not discriminate an instantaneous and a durative event in this case, just regard them as events anchored on certain day.

For an event which last more than one “day”, such as several “day”, “week”, “month”, “year” or “century”, if two boundaries of the interval can be identified from the text, events are represented by a set of dots, i.e. one dot per day in the mentioned intervals (see Figure 5.3, the lower part). For example, if we know “Peter arrived at Hong Kong in July, 2005”, we can use a set of dots to denote the event with one dot per day in July 2005. We assume each mention of an event is of same importance, no matter an event is instantaneous or durative, and no matter how long the event lasts. Therefore the weight of one dot for the

previous example is thus equally distributed to 30 day, i.e. 1/30, and the weight of the event is also equal to 1.

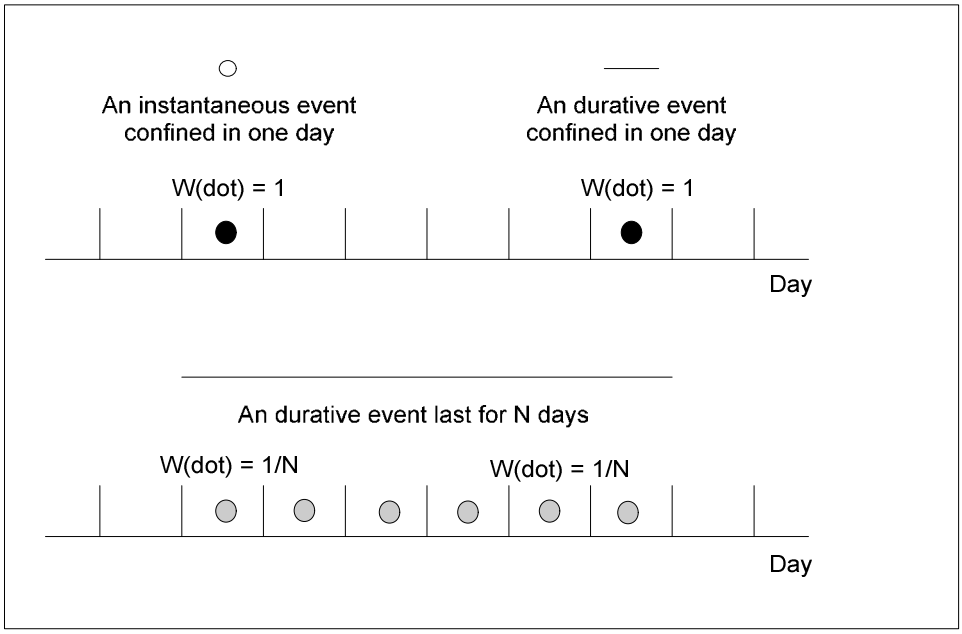

Figure 5.3 Representation for events with two boundaries

5.2.3 Representation of Event with One/Zero Boundary

Some events may have incomplete description about their boundaries on the time line. The time interval of the event in example (1) starts from the reference time “next Tuesday”, but there is no description about the end of the event. Similarly, the event in example (2) end at the reference time “2001” and has no beginning time. There is no beginning time and ending time about the event in example (3).

(1) Tom will leave New York after next Tuesday.

(2) John lived at Australia before 2001.

(3) Smith stays at New York now.

Although one or two boundaries of the events of these types can not be located on the time line, people always refer a time interval which is adjacent to the reference time. If the reference times are not mentioned clearly in sentences, such as example (3), the speech times or publication dates are used instead. For an instantaneous event, it may occur at a time point

in this interval. For a durative event, it may last the whole period. For both instantaneous and durative event, we split the event weight to this time interval. For these events, we formulate them with a series of discrete dots on the time line around or besides the reference time (see Figure 5.4). For an instantaneous event, a dot represents the possibility that the event occur in this temporal unit. For a durative event, a dot represents a part of the event weight.

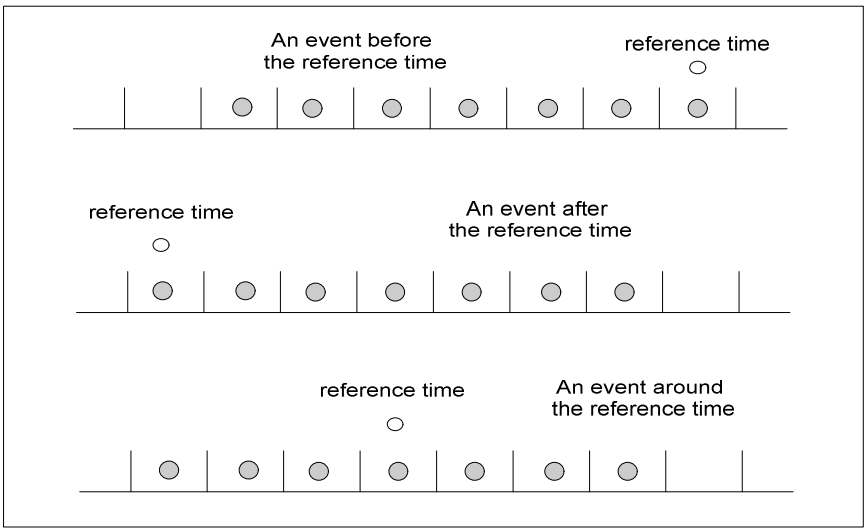

Figure 5.4 Representation for events with one/zero boundary

Normally an event time should be confined within a scope, but in previous examples one or two boundaries of the events are not known. In example (1), we assume Tom will leave New York in several days after next Tuesday. In example (2), we assume John lived at Australia for several years and may left there at 2001. In example (3), we assume Smith arrived at New York several days before and he will leave there several days later. From these assumptions, it can be seen that the unlimited time interval of an event is replaced by a limited one which is near the reference time. Therefore the unknown boundaries can be guessed, i.e., we have to decide how long the distance is between a boundary and the reference time.

The following question is how many dots should be employed for the events with one or zero boundary? Based on the observation that events in news reports commonly occur near the reference time within a week, we tentatively employ 7 dots to denote the distance and place them into 7 temporal slots besides the reference time with same granularity, each dot

per slot. If a reference time is given with the temporal unit “year”, “week” or “day”, then the unit of time slots is also “year”, “week” or “day”.

For an event before or after the reference time, 7 dots are placed on 7 time slots which are immediately before or after the reference time. For the event which occurs around the reference time, 3 dots are inserted before the reference time and 3 dots after the reference time. 1 dot is inserted on the day of the reference time. Note that sum of weights of 7 dots should be 1. They are shown in Figure 5.4.

It is assumed that the events are more likely to occur at the time slots closer to the reference time. We assume normal distributions of the weights for the events that occur around the reference time.

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

In the above equation, μ is the reference time, 3σ is equal to 7. By adding the right part of the symmetrical distribution function to the left part, we get the distribution function for the events occurred before the reference time (Figure 5.5, “before”). Similarly, we can get the distribution function for the events occurred after the reference time (Figure 5.5, “after”). The weight of each dot D on certain time slot x, i.e. $W(D, x)$, in Figure 5.4 is equal to the area which is under the corresponding distribution function within the scope of the corresponding time slots in Figure 5.5.

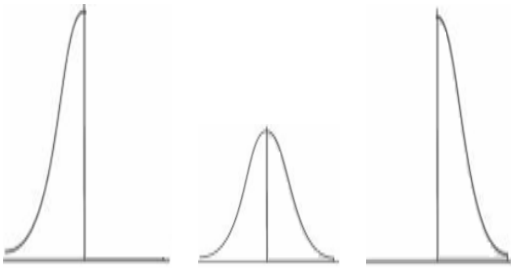

Figure 5.5 Distribution functions for weight of dots in Figure 5.4

Once $W(D, x)$ is calculated, if the temporal unit is coarser than “day” (1 unit = y days), the weight of the event term/element on each day is computed as $W(D, x)/y$. Otherwise, if the

temporal unit is finer than “day” (1 day = y units), the weight of the event term/element on each day is computed as

$$\sum_{t=x-(y/2)}^{t=x+(y/2)} W(D,t)$$

5.3 Event Weighting and Sentence Selection

In Section 5.2, each event term or event element occurrence can be represented by a dot on a day or dots on a series of days. The distributions of two sample event terms or event elements on the time line are illustrated in Figure 5.6. The weight of each dot on the time line can be different. Because of the reason we explained in Section 5.2, temporal unit “day” is employed to measure the time line.

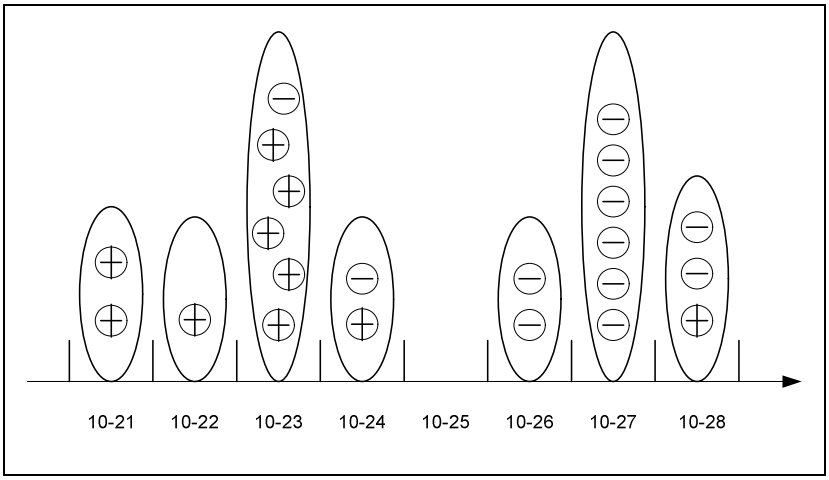

Figure 5.6 Two event terms/elements on the time line (⊕ : an event term/element. ⊖ : another event term/element)

5.3.1 Event Weighting Schemes

Two weighting schemes, *tf***idf* and χ^2, are used to measure the importance of an event terms/elements on a day. Here, *tf* is the sum of weights of the instances of an event term/element on a day and *idf* is equal to 1 over the number of days on which the event term/element happened. Multiple instances of the same event term or event element in one day may exist. The algorithm is similar to that in [Swan and Allan 2000]. It is defined as:

	E	$\neg E$
$t = t_0$	a	b
$t \neq t_0$	c	d

$$\chi^2 = \frac{N(ad - bc)^2}{(a+b)(a+c)(b+c)(b+d)}$$

In the above equation, ***a*** is the sum of the dot weights on the day t_0, and all these dots should be associated to the same event term/element *E*. ***b*** is the sum of the dot weights of the event terms/elements other than *E* on the same day t_0. ***c*** is the sum of the dot weights on the days which are not t_0, and these dots represent the same event term/element *E*. ***d*** represent the sum of the dot weights on the days other than t_0, and the dots represent the event terms/elements which are not *E*. Since *N* is a constant, it does not influence sentence ordering. We omit it in our computation. Then the weight of a sentence is achieved by summing up all the weights of event terms /elements contained in the sentence. The weight of each event term/element is the sum of weights of the dots which are associated to this event term/element.

5.3.2 Sentence Selection Strategies

We use two strategies for sentence selection, i.e. sequential selection and round robin selection. For both of them, sentences are sorted according to their weights first. Sequential selection selects the sentences one by one in each loop, until the length limitation of the summary is reached. Round robin selection selects the sentences with the highest score for each year, month, or day in each loop until the length of the summary is up to the limitation. We refer to these three round robin approaches as Robin_Year, Robin_Month, and Robin_Day. We are aware of some techniques, like MMR [Carbonell and Goldstein 1998], have been successfully used to handle the problem of redundancy. The focus of this study is to study how to formulate event temporal distribution and how to exploit it to select the most important sentences, so these techniques are not emphasized.

5.4 Experiment and Evaluation

5.4.1 Data Set and Evaluation Methods

DUC 2001 provides 30 English document clusters and the corresponding model summaries with up to 200 words. This data set is used in the experiments of Chapter 3. After our observation, we find that 10 clusters of the data set contain descriptions about topics shifting over time. It can be seen that a substantial percentage of clusters are highly relevant to time. Table 5.1 describes the topic of each cluster and give whether the topic is temporal or non-temporal. For example, Cluster D41 reports fires in USA at 1926, 1977, 1985, 1987, 1988, 1989, 1990 and 1991. Figure 5.7 and Figure 5.8 give paragraphs extracted from source documents and the model summary. We can see the model summary consists of descriptions about fires at different years. The italic paragraphs of the model summary in Figure 5.8 come from the source paragraphs in Figure 5.7.

We first evaluate the proposed approaches on two temporal evolving clusters only. The event-based summarization approaches without time perform worst on these two clusters. Encouraged by the results, then we extend the evaluation on 10 temporal evolving clusters and conduct more detailed experiments. ROUGE is used to evaluate the quality of machine generated summaries by comparing them with the model summaries according to their word unigram overlap, bi-gram overlap or overlap with long distance. It is an automatic evaluation tool used widely in DUC conferences, but the limitation is that it can not compare meaning overlap between our summaries and model summaries. Therefore we also invite a graduate to evaluate the system output according to meaning relevance to model summaries. We think the evaluation by a subject is more believable, although the procedure is time-consuming.

Cluster	Topic(s) in the cluster	Temporal
d04	Huge damage in U.S. by hurricane Andrew	No
d06	Racism and brutality against minorities in U.S. police forces	No
d11	Introduction about source, phenomenon and prevention of tornadoes	No
d12	The welfare reforms conducted by presidents Reagan and Clinton	No
d13	Introduction to a Supreme Court justice, Clarence Thomas	No
d14	Airplanes, locations and procedures of military aircraft crashes	No
d15	Introduction to tuberculosis, a kind of disease	No
d19	Different attitudes to illegal aliens in U.S.	No
d22	The damage of forest fires in U.S.	No
d27	Discussion about gun control in U.S.	No
d34	Introduction to hurricanes and the related research	No
d31	Illegal use of steroid by Canadian Ben Johnson at Seoul Olympics	No
d39	Construction of the tunnel under English Channel	No
d43	The diamond business of De Beers Consolidated Mines Ltd	No
d44	Argument about the North American free trade agreement	No
d53	Introduction to Shining Path guerrillas in Peru	No
d54	Argument about term limitations on elected officials	No
d56	Introduction to diabetes in U.S.	No
d57	Introduction to earthquake and related research	No
d59	Descriptions about airplane crashes and the reasons.	No
d05	Development of mad cow disease (1986, 1988, 1993 and 1994)	Yes
d08	Events of solar eclipse (1868, 1919, 1988, 1990 and 1991)	Yes
d24	The life of Elizabeth Taylor	Yes
d28	Events of marathon racing in different cities (1989, 1990 and 1991)	Yes
d30	The Third Word debt (1984, 1987, 1990, 1991 and 1993)	Yes
d32	Events about the aground ship of Exxon (1989 and 1990)	Yes
d37	Assassination events (1989, 1990, 1991 and 1994)	Yes
d41	The wildfires in U.S. (1977, 1985, 1987, 1988, 1989, 1990 and 1991)	Yes
d45	Slovenia and Serbia (1988, 1989,1990, 1991, 1992 and 1994)	Yes
d50	Events about drought in U.S. (1988, 1989 and 1990)	Yes

Table 5.1 Topics of clusters in DUC 2001 data set

Document ID: AP890805-0126

As nearly half the acreage a fire burns within 150 miles, specialists in the logistics centre marshal resources from around the nation. On Saturday, there were about 220,000 acres ablaze in four states, with 102,000 of them in Idaho and the rest in Oregon, California and Utah.

Document ID: AP890801-0025

Fire has charred more than 1.3 million acres of forest and range land since January in the contiguous United States, compared to 2.1 million acres by this time last year, fire officials said Monday.

Document ID: LA081490-0030

The Yosemite fires, which have scorched more than 15,000 acres and are still out of control, and the recent Santa Barbara and Glendale fires, which destroyed nearly 500 homes, are only the largest of hundreds of blazes in the state. "We've got lots of summer ahead of us and we've already burned 600 or 700 structures," said Deputy Chief Keith Metcalfe of the state's southern regional fire fighting crew in Riverside.

Document ID: SJMN91-06071022

That approach helped last year in the Umatilla National Forest in north-eastern Oregon, where 170 fires were reported, said Gordon Reinhart, a fire and recreation officer with the U.S.

Figure 5.7 Paragraphs extracted from source documents in Cluster d41

Large Western wildfires occurred in the 1700 s and probably at 200 - 300 year intervals over the past 10,000 years.

In 1926, wildfires destroyed 28 million acres.

1977's Sycamore Canyon fire destroyed 200 homes near Santa Barbara.

1985's Los Altos Hills fire destroyed 100 acres and 12 homes, and a six - day fire above Lexington Reservoir destroyed 14,000 acres and 42 homes.

In 1987, 900,000 acres burned in California. 1,500 fires attacked Klamath National Forest in one month.

1988 was the most expensive fire season in US history, with 25,000 fire-fighters called in and $583.8 million spent to fight 75,000 fires that burned 5.9 million acres. 706,000 of Yellowstone's 2.2 million acres burned over four months, stopped by light rain and snow in September. Smoke sickened 12,000 fire-fighters. The "let - burn" policy was lifted in July.

From January - July 1989, 1.3 million acres had burned in the contiguous US . In August ,102,000 acres in Idaho and 118,000 in Oregon , California and Utah were ablaze .

In 1990, fires in Santa Barbara destroyed 600 homes and did $ 200 million damage, and burned 15,000 acres in Yosemite. 170 fires burned in Oregon's Umatilla National Forest.

In October 1991, fire devastated the Oakland hills. Nine Saratoga foothills fires were deliberately set.

Figure 5.8 The model summary of Cluster d41

5.4.2 Preliminary Evaluation on Two Clusters

First, we investigate two event weighting schemes *tf*idf* and χ^2. Sentences are selected sequentially according to their weight ranking. The baseline is the instance-based event summarization without considering temporal distributions. The influence of context on event term is not considered in this base line. The results given in Table 5.2 shows that the event-based summarization with temporal *tf*idf* weighting scheme outperforms both the baseline and χ^2 significantly.

	Baseline	*tf*idf*	x^2
ROUGE-1	0.271	0.322 (+ 18.8%)	0.282 (+ 4.1%)
ROUGE-2	0.029	0.081 (+179.3%)	0.024 (-17.2%)
ROUGE-W	0.101	0.120 (+ 18.8%)	0.104 (+ 3.0%)

Table 5.2 Temporal-oriented event summarization with *tf*idf* and χ^2 weighting scheme (sequential sentence selection)

Base on the temporal-oriented summarization with *tf*idf* weighting scheme, the sequential sentence selection and round robin selections are compared in Table 5.3. The sequential selection performs best.

	Robin_Year	Robin_Month	Robin_Day	Sequential
ROUGE-1	0.268	0.290	0.314	0.322
ROUGE-2	0.018	0.018	0.057	0.081
ROUGE-W	0.093	0.105	0.123	0.120

Table 5.3 Temporal-oriented event summarization with sequential and round robin sentence selection (*tf*idf* weighting scheme)

5.4.3 Further Evaluation on Ten Clusters

In the following sets of experiments, the baseline and temporal-oriented event summarization with *tf*idf* weighting scheme are compared on ten clusters with different evaluation criteria. First, word-based ROUGE evaluation compares the overlaps between

words in model summaries, event-based summaries and temporal-based summaries. As word-based ROUGE is incapable to tell whether the events or meanings of two summaries are same or relevant, we then consider event-based ROUGE and a subjective evaluation.

The event-based ROUGE evaluations including instance-based and concept-based ones compare overlaps of event instances and concepts in the summaries. To examine the meaning overlaps, we invite a graduate to judge whether each sentence in system generated summaries is semantically relevant to any sentence in model summaries. The experiments are described as follows.

5.4.3.1 ROUGE Evaluation with Words

ROUGE-1 scores of event-based and temporal-based summaries are compared in Figure 5.9. The average ROUGE-1 of them is 0.326 and 0.310 respectively. Temporal summaries outperform event-based summaries in seven of ten clusters, but the average score of temporal summaries in overall is lower than that of event-based summaries.

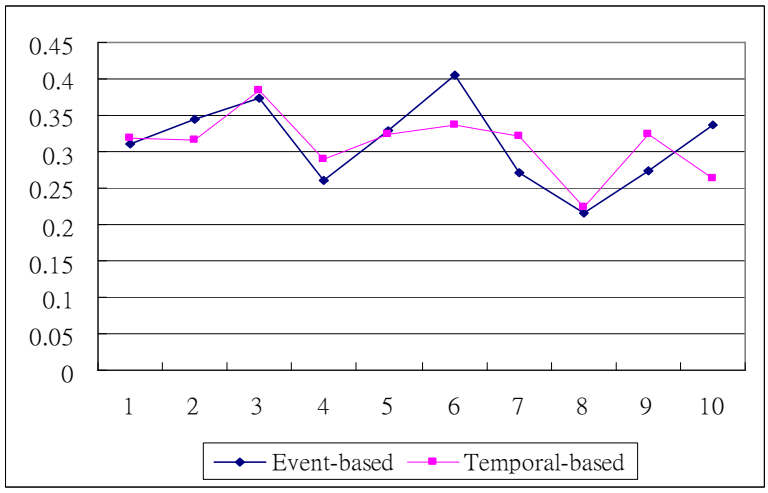

Figure 5.9 Evaluation results on overlaps of words

5.4.3.2 ROUGE Evaluation with Events

Event terms and event elements are extracted from model summaries, event-based and temporal-based summaries in the same way, as they are extracted from original documents.

We call each event occurrence as an event instance. The instances of the same event are merged into an event concept. Then, ROUGE is run based on these instances and concepts.

Figures 5.10 and 5.11 compare the ROUGE-1 scores of event-based summaries and temporal-based summaries based on event instance and concept respectively. The average ROUGE-1 scores of event-based summarization are 0.160 and 0.157 respectively. The average ROUGE-1 scores of temporal-based summarization are 0.144 and 0.145 respectively. Taking a closer look at Figures 5.10 and 5.11 we find that temporal summaries outperform event-based summaries in five clusters based on either instances or concept.

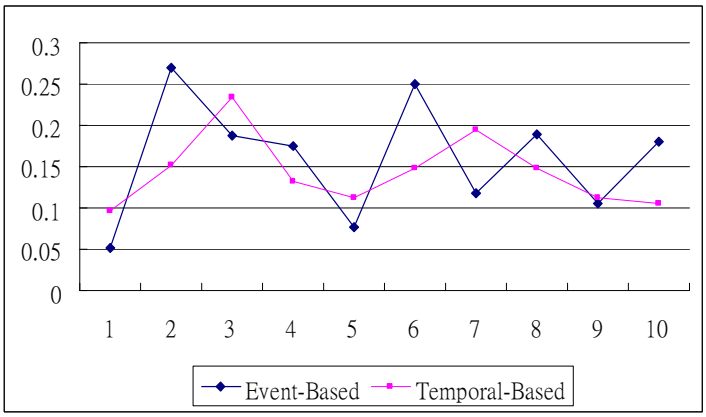

Figure 5.10 Evaluation results on overlaps of event instances

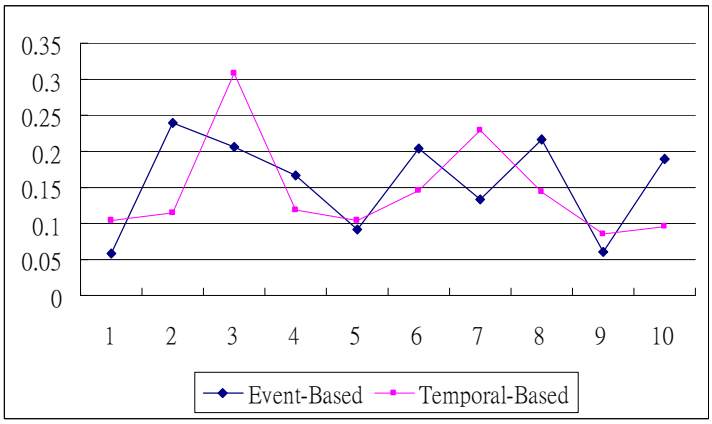

Figure 5.11 Evaluation results on overlaps of event concepts

5.4.3.3 Subjective Evaluation

The evaluation guideline is described as follows. First a subject is asked to read all the documents and model summaries. Given a machine generated summary, the subject judges whether each sentence in it is relevant to any sentence of the corresponding model summary.

If meanings (participants, action, when, where, method, status) of a model sentence are same as the meanings of a machine generated sentence, or the subject can infer all the meanings of the former from the meanings of the later, the machine generated sentence receives a score "1". If the subject can just infer part of the meanings of a model sentence from the machine generated sentence, then the later receives a score "0.5". If the subject can not infer any meanings of any model sentence from a machine generated sentence, then the later is given "0". If multiple rules can be applied to a sentence, then its score will be the maximum.

After assigning a score to each sentence in machine generated summaries, we sum up the scores of all sentences in it as the score of the summary. The results are shown in Table 5.4. It can be seen that temporal-based summarization get significant improvement.

	C05	C08	C24	C28	C30	C32	C37	C41	C45	C50	Ave.
Event	1.5	2.5	4.0	0	0	2.0	1.0	0	1.5	2.0	1.5
Temp	3.0	3.5	7.0	1.5	2.5	3.0	3.5	0	4.0	2.0	3.0

Table 5.4 Evaluation results given by a subject

5.4.4 Evaluation with Auto-Tagged Temporal Information

We evaluate our summarization approaches with manually tagged temporal information in previous experiments. In this section we evaluate the best summarization approach, which consists of temporal *tf***idf* weighting scheme and sequential selection strategy, on automatically tagged data. The ten clusters used in experiments of Section 5.4.3 are employed as test data again. Fist temporal expressions in these clusters are extracted and normalized by our English system described in Section 4. Then heuristics are employed to assign temporal values to corresponding clauses. The evaluation results of the summarization approach with manually tagged and auto-tagged temporal information are shown in Table 5.5. It can be seen that the summarization performance with auto-tagged temporal information is lower than that with manually-tagged information, but it is higher than that of baseline (Table 5.4). This experiment reflects that our temporal information processing system is reliable. For comparison, we conduct our temporal-oriented event-based

summarization approach (temporal *tf*idf* and sequential selection) on other 20 clusters, which do not contain evolving events on the time line. The average relevant sentence in summaries is 1.3. It can be seen that our temporal-oriented summarization is suitable for those temporal related clusters and not suitable for other types of clusters.

	C05	C08	C24	C28	C30	C32	C37	C41	C45	C50	Ave.
Auto	1.5	4.0	5.5	1.0	1.0	2.5	3.0	0	2.0	2.0	2.3
Manual	3.0	3.5	7.0	1.5	2.5	3.0	3.5	0	4.0	2.0	3.0

Table 5.5 Evaluation results with auto-tagged and manually-tagged temporal information

5.5 Discussion

Why the improvement of temporal summaries is visible in subjective evaluation? Look at some basic statistics first. In average, the number of sentences in a temporal summary and an event-based summary is 6.6 and 5.7 respectively. The average numbers of event terms and event elements in an event-based summary and temporal-based summary are very close, i.e. 48 and 49 respectively. Therefore more sentences are included in temporal summary and more information other than the extracted events might be included.

Next, we examine the temporal distribution of the events selected into the temporal-based summaries to see whether the selected events are most important. The model summaries of these ten clusters consist of a sequence of descriptions about a same or similar topic over time, such as a film star, assassinations, etc. We select the sentences which burst on particular periods. Events in these sentences are found mentioned frequently on bust periods but seldom on the other periods. The burst sentences are more likely the focus of burst periods. Therefore, they are more likely to be relevant with model sentences.

Then, we examine the importance of events selected in event-based summaries. Event-based summarization without considering temporal distribution selects the sentences which contain event terms and elements with higher centroid scores. The centroid scores are the average *tf*idf* weights over all documents. The events selected in this way are the "centroids" of clusters, but they may not be focuses at different periods. Therefore, these

sentences are less likely included in model summaries, compared with the sentences selected by temporal-based summarization. For example, the distribution of Cluster d37 summary sentences is presented in Figure 5.12. The dark, gray and white dots denote the events which are relevant, partially relevant and irrelevant to the events in model summaries. It is easy to see in Figure 5.12 that temporal-based summaries are better than event-based summaries though they are still not completely matched with the events in model summary.

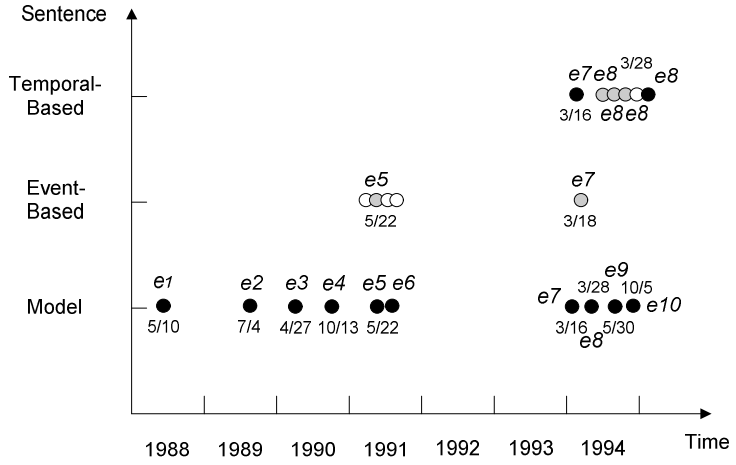

Figure 5.12 Distribution of events in summaries on the time line

As discussed in Section 5.3, temporal-based summarization assigns weights to sentences based on event distribution on time line. The final summaries may include the most important events in certain periods but may not represent the whole trends well. To overcome this drawback, we apply clustering technologies to group events into different time periods. Sentences with the highest weights in each group are selected.

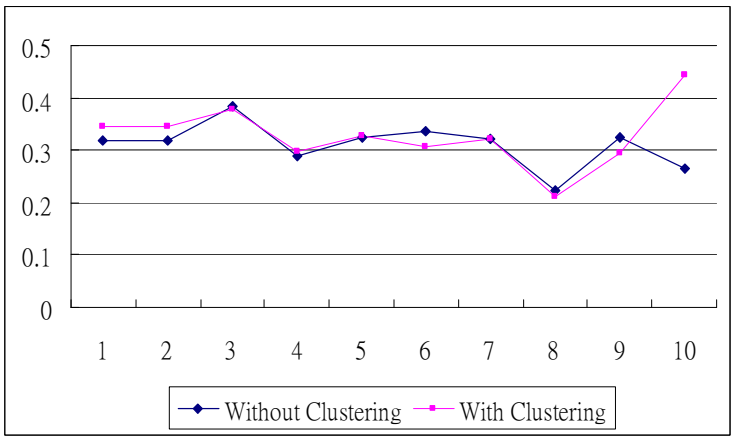

Figure 5.13 Temporal-based summarizations with and without sentence clustering

DBSCAN, a density-based clustering approach, is employed. It has no need to set the number of clusters and terminal conditions in advance. It has only two configuration parameters, Eps and MinPts. Eps means the maximum radius of the neighbourhood, and MinPts denotes the minimum number of points in an Eps-neighbourhood of that point. Eps and MinPts are tuned experimentally. The optimal combination is (7, 5) in our experiments. The results of *tf***idf* weighting scheme with sequential and clustering-based sentence extraction strategies are shown in Figure 5.13. ROUGE-1 score is improved from 0.325 to 0.310 by 4.8% when using clustering.

5.6 Chapter Summary

In this study, whether integrating temporal information can enhance event-based summarization is investigated. *tf***idf* and χ^2 weighting schemes are employed to evaluate importance of event term/element on each day. Three sentence selection strategies are compared: sequential, round robin and clustering selection. Experiments show that *tf***idf* weighting scheme based on event temporal distribution performs better than event-based summarization without consideration of time. It can be seen that event temporal distribution is helpful to summarize the trend of news topics and clustering selection can improve the quality of summaries under the *tf***idf* weighting scheme.

We have three plans for the near future. First, we will investigate how and to what extend event terms and elements contribute to temporal-based summarization. Second, we would like to see how the redundant sentences can be effectively removed or compressed in order to allow more relevant sentences included in summaries. Finally, sentence order and discourse structure will be investigated for generating more coherent summaries.

Chapter 6

Integration of Summarization Features under Learning-Based Framework

6.1 Chapter Overview

Extractive summarization approaches select important sentences into summaries by exploiting different types of features. For example, length of sentences and position of sentences are used as two features in MEAD, a typical extractive summarization system [Radev et. al 2004]. Typical sentence weighting schemes might use a linear function which sums the weighted values of different features. The drawback of such schemes is that weight parameters have to be tuned experimentally. This is time-consuming as it requires re-running experiments for different styles of documents, and it can inadvertently overlook or exclude the best combinations of parameters. One possible approach to this problem is to use learning-based classification to make the weighting scheme optimal [Kupiec et al. 1995; Conroy and Schlesinger 2004]. Successful results reported in the literature motivate us to improve this framework by investigating more effective summarization features.

The performance of learning based approaches depends heavily on input features. Recently different sentence features are explored individually, but most of them are not incorporated into a learning based approach and their functions are not well studied when they are combined together. Word relevance in a sentence and relevance between sentences are not exploited in previous learning based approaches. Event features of sentences are also

neglected in learning based approaches. Therefore we conduct this study on learning-based framework and various sentence features, especially content, relevance and event features.

The framework in this study [Wu et al. 2007b] classifies importance of sentences based on four types of features: surface, content, event and relevance. Surface features include, for example, the number of words in a sentence or its position in a document. Content features would, for example, refer to the use of high frequency words. Event features refer to importance of events contained in sentences. Finally, relevance features refer to those less easily observed features of a sentence, such as the relevance between a sentence and the first sentence of a paragraph, which contribute to the coherence of a document.

The process by which sentences are identified as "important" and then included in a summary is as follows. First, each sentence is audited for the four types of features. Using these features, a learning-based classification is then used to classify the sentences as unimportant and important. After the classification, the testing sentences are re-ranked and assigned a weight score. Important sentences are included into summaries first, then unimportant sentences if the length limitation of summaries is not reached. When important or unimportant sentences being extracted, sentences with the higher re-ranking scores are considered first. Experiments show that the proposed framework achieves competitive results and relevance features are able to improve the summarization performance obviously.

The remainder of Chapter 6 is organized as follows. Section 6.2 presents learning-based extractive summarization. Section 6.3 describes experimental setup and evaluations. Section 6.4 discusses the results and Section 6.5 summarizes the paper.

6.2 Learning-Based Extractive Summarization

We propose a learning-based framework for extractive summarization. It contains three parts: the classification model, features employed in the model and sentence re-ranking algorithms. The summarization procedure is described as follows. First each sentence is extracted from documents and different types of features are checked. Then values of the features are sent to

a classification model and the importance of the sentence is given by the model. After the importance of sentences is adjusted by a re-ranking model, the most important sentences are used to form the final summary. Figure 6.1 illustrates our learning-based extractive summarization

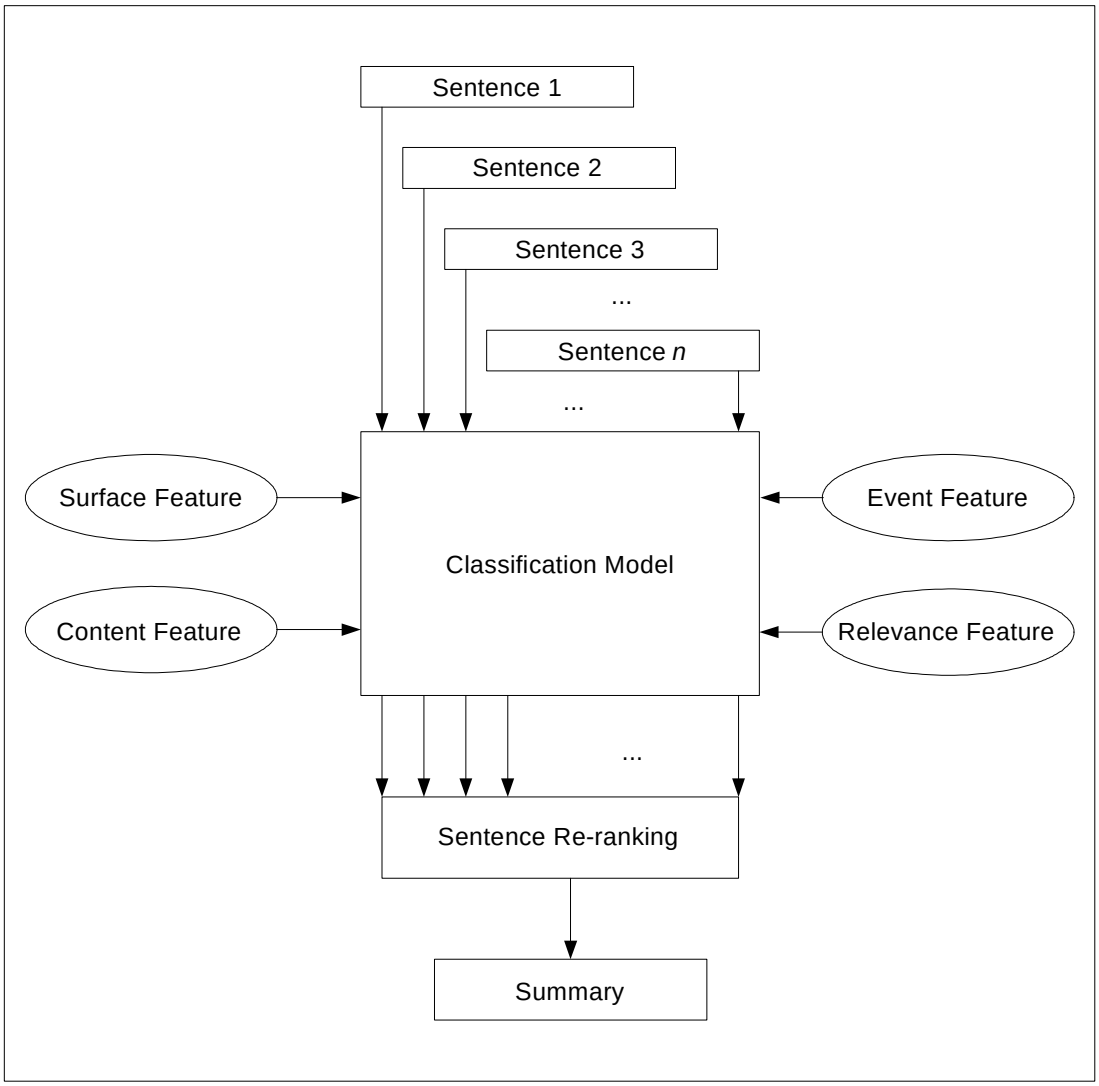

Figure 6.1 A learning-based summarization framework

6.2.1 Classification Model

The task of extractive summarization is to select important sentences from documents. In this study, we consider the evaluation on sentence importance as a classification task. Sentence importance can be measured by scaled values, ranging from 0 to 1 and indicating

importance extent, or it can be measured by binary values, indicating whether the sentence concerned is important or not.

Categories involved in classification can be two, e.g. important and unimportant, or multiple, such as very important, important, less important and unimportant. However, it is difficult to clearly define the boundaries of multiple categories and sparse data may bring another problem. Therefore we choose binary classification. Some classifiers are capable of conducting binary classification with probability estimates, i.e., how likely a sentence is important. We have experimented with probabilistic classification. It does not perform as well as binary classification, so it is not adopted.

To train the binary classifier, a labeled sentence bank is constructed semi-automatically. First, an automatic content similarity evaluation tool, ROUGE, is adopted to filter out sentences in documents, whose wordings are quite different from the sentences in model summary. The filtered sentences are labeled unimportant automatically. The remaining sentences are then manually labeled as important if their contents overlap the contents of any model summary sentence, or unimportant if not. The important sentences are used as positive training examples. All the other sentences are kept as negative examples.

Considering SVM classifiers have achieved promising performance in many applications, we choose one of them, LIBSVM [Chang and Lin 2001] to carry out the classification task introduced before. Given training data, the classifier train a classification model for later testing. In the procedure of testing, the input of the classifier is a feature vector and the output is the label for each classification object. For sentence classification, the features can be length or position of sentence, etc. The label of sentences can be important or unimportant. Actually, other binary classifiers are also applicable for this application.

6.2.2 Features for Classifications

This section provides a detailed description about four types of sentence features to be used in this summarization schemes: surface, content, event, and relevance.

6.2.2.1 Surface Features

In this study, surface features are the number of words in a sentence, sentence position in the document, and the number of quoted words (see Table 6.1).

Name	Description
Position	1/sentence No. (Real Number)
Doc_First	whether it is the first sentence of a document (Binary Number)
Para_First	whether it is the first sentence of a paragraph (Binary Number)
Length	The number of words in a sentence (Real Number)
Quote	The number of quoted words in a sentence (Real Number)

Table 6.1 Surface features

US cities along the Gulf of Mexico from Alabama to eastern Texas were on storm watch last night as Hurricane Andrew headed west after sweeping across southern Florida, causing at least eight deaths and severe property damage. The hurricane was one of the fiercest in the US in decades and the first to hit Miami directly in a quarter of a century..

……

Andrew, the first Caribbean hurricane of the season, hit the eastern coast of Florida early yesterday, gusting up to 165mph. It ripped roofs off houses, smashed cars and trucks, snapped power lines and uprooted trees before heading out over the Gulf.

Figure 6.2 Examples of surface features[1]

We assume that (1) the first sentence in a document or in a paragraph is important; (2) the sentences in earlier parts of a document is more important than the sentences in later parts; (3) a sentence is important if the number of the words (except stop words) in it is within a certain range; (4) the sentence containing many quoted words is unimportant. For example, "he said '……' " is not likely to be included in a summary. Some examples are illustrated in Figure 6.2.

[1] The sentence marked with "…", "…" or "…" is the first sentence in a document, in a paragraph, or at the beginning of a document.

6.2.2.2 Content Features

In this work, we make use of three different, recognized definitions of content-bearing words i.e., as centroid words, signature terms, and high frequency words.

Centroid words

Given N documents, each word t is weighted as the average of *tf*idf* scores across the documents. The formula is

$$w(t) = \frac{\sum_{i=1}^{N} tf_{ti} * idf_{t}}{N}$$

where $w(t)$ is the weight of word t. tf_{ti} is the frequency of t in the ith document, idf_t is the inverse document frequency of t. The *idf* factors are computed based on TDT corpus in [Radev et al., 2004]. However, the inverse document frequency, *idf*, may be computed in variant ways. Section 6.3 presents an experimental comparison of two ways of computing *idf*, from the whole DUC2001 document set or from a single document set. Centroid approach with the latter *idf* calculation is named CentroidVar in this paper.

Signature terms

The signature term approach assumes that documents in a document cluster (i.e. a set of relevant documents) are regarded relevant to its topic, while documents in other clusters are not relevant to the topic. Based on the distribution across relevant and un-relevant documents, signature terms are extracted according to likelihood ratio λ [Dunning 1993].

$$-2\log\lambda = -2\log\frac{b(O_{11}; O_{11}+O_{12}, p)\times b(O_{21}; O_{21}+O_{22}, p)}{b(O_{11}; O_{11}+O_{12}, p_1)\times b(O_{21}; O_{21}+O_{22}, p_2)}$$

$$b(k; n, x) = \binom{n}{x} x^{k} (1-x)^{(n-k)}$$

where O_{11} and O_{12} denote the frequencies of term t_i occurring in the relevant set and irrelevant sets respectively, O_{21} and O_{22} denote the frequencies of terms $\neg t_i$ occurring in the relevant set and irrelevant sets respectively. Terms whose $-2\log\lambda$ is higher than a threshold are extracted as signature terms.

High frequency words

It is quite encouraging that using high frequency words only can give promising results [Nenkova et al., 2006]. High frequency words are also exploited in our feature-based classification as a content feature. The threshold of word frequency is tuned experimentally.

CentroidVar_Uni	CentroidVar_Bi
mr 1.92	bush mr 2.22
lloyd 1.25	mr president 1.67
miles 1.25	clinton mr 1.67
political 1.11	mr mr 1.39
city 1.07	losses pounds 1.39
losses 1.06	cent insurance 1.39
gulf 0.97	mr pounds 1.39
catastrophe 0.89	pounds uk 1.11
uk 0.89	house mr 1.11
president 0.88	pounds royal 1.11

Figure 6.3 Top ten uni-grams and bi-grams of centroid words

SigTerm_Uni	SigTerm_Bi
andrew 204.78	dollars hurricane 109.85
florida 180.12	dollars losses 93.40
losses 146.67	damage dollars 82.55
dollars 144.56	hurricane yesterday 77.01
insurer 142.26	florida hurricane 73.15
insured 142.26	caused dollars 69.73
insurance 142.26	damage hurricane 62.70
insurers 142.26	florida storm 60.70
louisiana 125.31	dollars total 55.31
yesterday 107.50	1989 dollars 52.58

Figure 6.4 Top ten uni-grams and bi-grams of signature terms

FreqWord_Uni	FreqWord_Bi
hurricane 35.00	andrew hurricane 17.00
florida 34.00	dollars hurricane 14.00
dollars 33.00	andrew florida 13.00
andrew 27.00	florida hurricane 13.00
mr 24.00	claims insurance 13.00
insurance 23.00	dollars losses 12.00
losses 22.00	damage dollars 12.00
yesterday 22.00	damage hurricane 10.00
louisiana 20.00	hurricane yesterday 10.00
damage 20.00	caused dollars 10.00

Figure 6.5 Top ten uni-grams and bi-grams of high frequency words

While previous work concerns only uni-grams of centroid and high frequency word, we believe that bi-grams present more precise concepts by considering word intra-sentence

relevance. Therefore we extend the proposed approaches from uni-grams to bi-grams. In our study, bi-grams are soft rather than rigid patterns. In other words, not only two adjacent words, but also two long-distance words within a sentence are considered. The order of the two words is ignored in order to cope with the flexibility of language. For example, "hurricane Andrew" is same as "Andrew hurricane". The top ten uni-grams and bi-grams examples from centroid words, signature terms and high frequency words are shown in Figures 6.3, Figure 6.4 and Figure 6.5 respectively.

Table 6.2 summarizes the content features we investigate. Content features can be measured by either the sum or the number of weights of uni-grams or bi-grams the sentence contains.

Name	Description
CentroidVar_Uni	The sum (or number) of the weights of centroid uni-gram (Real Number)
CentroidVar_Bi	The sum (or number) of the weights of centroid bi-grams (Real Number)
SigTerm_Uni	The sum (or number) of signature uni-grams (Real Number)
SigTerm_Bi	The sum (or number)of signature bi-grams (Real Number)
FreqWord_Uni	The sum (or number) of the weights of high frequency uni-grams (Real Number)
FreqWord_Bi	The sum (or number) of the weights of high frequency bi-grams (Real Number)

Table 6.2 Content features

6.2.2.3 Event Features

As event is a natural unit to represent meanings about a topic, we have investigated event-based summarization approaches [Wu 2006; Li et al. 2006b] and achieved promising results. To incorporate event features, instance-based and concept-based summarization approaches are employed in this study. The two approaches give two weight scores for each sentence and the two scores are used as two event features. They are shown in Table 6.3.

Name	Description
Event_Instance	A weight score given by instance-based event summarization (Real Number)
Event_Concept	A weight score given by concept-based event summarization (Real Number)

Table 6.3 Event features

6.2.2.4 Relevance Features

Relevance features are incorporated to reflect inter-sentence relationships. It is reasonable to believe that the sentences relevant to important sentences or many other sentences may provide important concepts or serve as the pivots of related concepts. Based on the assumption that the first sentences in document or paragraphs are most important in establishing the topic, all other sentences are then compared with these sentences for relevance. Sentence relevance is measured by comparing pairs of sentences using a word-based cosine similarity.

Name	Description
FirstRel_Doc	Similarity with the first sentence in the document (Real Number)
FirstRel_Para	Similarity with the first sentence in the paragraph (Real Number)
PageRankRel	PageRank value of the sentence based on the sentence map (Real Number)

Table 6.4 Relevance features

We also conduct pair wise sentence comparison with word-based cosine similarity. Two sentences are regarded relevant if their similarity is above a threshold. Based on the built sentence map, PageRank algorithm is applied to evaluate how important a sentence is. The three relevance features concerned are shown in Table 6.4.

6.2.3 Sentence Re-ranking

A problem associated with binary classification is that the numbers of the words in the important sentences identified by a classifier may not match the required summary length in

words. To overcome this problem, all sentences are re-ranked with a unified scheme. Sentences with higher priority are selected to generate summaries.

Different re-ranking algorithms are designed based on surface features or content features. They are evaluated and compared in Section 6.3.6. The following equation is an example of sentence ranking algorithm..

$$Rank_i = RankPos_i + RankLenght_i$$

where $RankPos_i$ is the rank of sentence i according to its relative position in a document (i.e. the sentence order No.) and $RankLength_i$ is rank of sentence i according to its length.

6.3 Experiment and Evaluation

6.3.1 Dataset and Evaluation Methods

Feature-based summarization is adaptive to various documents. We choose DUC 2001 document sets as the evaluation data. It contains 30 clusters of documents and a total of 308 documents. The numbers of the documents in a cluster is between 3 and 20. Each document cluster contains descriptions on a specific topic (e.g. Hurricane Andrew) and comes with 200-word model summaries created by NIST assessors. The task of our summarization approaches is to generate a 200-word summary for each cluster of documents.

As the purpose of extractive summarization is to select important sentences from source documents, it is similar to tasks of information retrieval. Precision [Salton and McGill 1983] can be employed to measure the percentage of true important sentences among all important sentences labeled by the classifier. Recall [Salton and McGill 1983] can be used to measure the percentage of true important sentences labeled by the classifier among all true important sentences. Precision and recall reflect the classification performance and they are based on sentence. When they are used as the evaluation standard, the limitation on summary is neglected. Actually the final summary may not include all extracted important sentences.

Manual and automatic evaluation methods can measure the quality of machine generated summaries. Since manual evaluation of summary qualities is time-consuming and may be subjective, the automatic evaluation package, ROUGE [Lin and Hovy 2003], has been widely used. ROUGE compares machine-generated summaries with model summaries based on uni-gram overlap, bi-gram overlap and overlap with long distance. It is a recall-based measure. ROUGE is not a comprehensive evaluation method. Instead, it provides a rough idea about how likely a machine-generated summary can be regarded as human-rephrased abstract. Generally speaking, a summary with a higher ROUGE score is better than another one with a lower score. We employ ROUGE in our following experiments.

6.3.2 Training Data Preparation

Twenty-five of the thirty document clusters are used as training data and the remaining five are used as testing data. There are 9212 training sentences and 1672 testing sentences in total. To label the training data, one may need to judge whether a sentence is relevant to any sentence in the model summaries. This strategy is quite time-consuming. So we introduce another approximate and cost bearable one. We use ROUGE to analyze the overlap between each sentence in a document cluster and three model summaries. Sentences are ranked according to its ROUGE-1 score. Only the sentences with higher ROUGE-1 scores are labeled by the subject.

An initial observation on a randomly selected document cluster, i.e. d04, is conducted to identify the appropriate threshold. There are 140 sentences in this cluster. Among the first 46 sentences of them, 26, i.e. more than half, are important. However, among the last 94 sentences, only 11 are important. If we manually label the first 46 sentences and automatically label the last 94 sentences as unimportant, 11 sentences will be labeled incorrectly. The error rate is just about 8%. In fact, a 200-word summary requires 5-8 sentences on average. Even for those 26 important sentences, not all of them can be included in the summary. On the other hand, 75% contents of model summaries can be covered by

those 26 important sentences. Based on these observations, we simply rank all sentences by ROUGE-1 scores first, and then manually label the top 1/3 sentences. The last 2/3 sentences are automatically labeled as unimportant.

6.3.3 Experiments on Individual Feature Groups

Features presented in Section 6.2.2 are evaluated by groups according to precision and recall. These two measures are employed to reflect the classification performance. If many true important sentences are selected, it can be inferred that features employed in the classification are effective. ROUGE is employed also to reflect the overlap between machine-generated summaries and model summaries. With the help of these two evaluation methods, we can see the performance about important sentence identification and the quality of final summaries.

LIBSVM is a suitable classifier for our extractive summarization. Each feature value sent to the classifier has been normalized as a real number which is between 0 and 1. As each feature value is a positive real number or zero, we set the normalized feature value as the original value over the maximum of this feature. A penalty parameter is necessary for the classifier to balance positive and negative training examples. In the following experiments, the recalls and precisions of different feature groups are given according to different penalty parameter values. Please note that there are about 36 important sentences in a cluster and the final summary contains about 5 to 6 sentences. Therefore the recall should be 0.15 or above. When this condition is satisfied, the parameter values which can bring highest precision are selected. For different features, there are different optimal parameter values.

6.3.3.1 Experiments on Surface Features

First each surface feature is evaluated separately to find the best one. The all of the five features are used. The precisions and recalls under different penalty parameter values are shown in table 6.5. From this table, we can see the most useful features are “pos” and “dsen1”, i.e. the position of a sentence in the document and whether a sentence is the first

sentence of a document. The combination of all surface features also gets the best performance, i.e. the precision score is 0.488 and the recall score is 0.146.

Parameter		1	2	3	4	5	6	7	8	9	10
len	P	0.000	0.000	0.550	**0.262**	0.188	0.194	0.191	0.182	0.169	0.151
	R	0.000	0.000	0.076	**0.153**	0.229	0.361	0.514	0.632	0.708	0.722
pos	P	0.000	**0.488**	0.488	0.488	0.276	0.276	0.231	0.197	0.197	0.169
	R	0.000	**0.146**	0.146	0.146	0.243	0.243	0.271	0.403	0.403	0.444
dsen1	P	0.000	**0.488**	0.488	0.488	0.488	0.488	0.488	0.488	0.488	0.086
	R	0.000	**0.146**	0.146	0.146	0.146	0.146	0.146	0.146	0.146	1.000
psen1	P	0.000	0.000	0.000	0.000	0.000	0.000	0.000	**0.099**	0.086	0.086
	R	0.000	0.000	0.000	0.000	0.000	0.000	0.000	**0.319**	1.000	1.000
mark	P	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	**0.090**	0.090
	R	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	**0.931**	0.931
All_Sur	P	0.000	**0.488**	0.477	0.392	0.291	0.258	0.231	0.224	0.211	0.206
	R	0.000	**0.146**	0.146	0.264	0.354	0.444	0.507	0.576	0.625	0.674

Table 6.5 The classification performance with surface features

6.3.3.2 Experiments on Content Features

We evaluate each content feature separately first, and then test groups of them. The purpose of these experiments is to compare different word-weighting algorithms, such as algorithms about centroid, signature term and high frequency word. The evaluation results on individual features and combination features are shown in Table 6.6 and Table 6.7 respectively. From these two tables, we find that the best individual feature is FreqWord_Uni, i.e. high frequency uni-gram. The best combination feature is the set of all six content features. The precision and recall brought by this feature combination are 0.407 and 0.167.

Parameter		1	2	3	4	5	6	7	8	9	10
SigTerm _Uni	P	0.000	0.333	**0.348**	0.327	0.321	0.244	0.216	0.219	0.201	0.180
	R	0.000	0.042	**0.167**	0.250	0.347	0.458	0.583	0.604	0.604	0.639
SigTerm _Bi	P	0.000	0.273	0.258	**0.275**	0.272	0.195	0.199	0.199	0.199	0.200
	R	0.000	0.021	0.056	**0.174**	0.215	0.264	0.299	0.299	0.299	0.313
CentroidVar _Uni	P	0.000	0.000	0.417	0.257	**0.243**	0.192	0.183	0.168	0.150	0.144
	R	0.000	0.000	0.035	0.063	**0.285**	0.431	0.750	0.819	0.847	0.875

CentroidVar _Bi	P	0.000	0.556	0.395	**0.304**	0.235	0.215	0.205	0.206	0.201	0.195
	R	0.000	0.035	0.104	**0.146**	0.278	0.368	0.417	0.486	0.514	0.556
FreqWord _Uni	P	0.000	0.316	**0.351**	0.333	0.295	0.248	0.189	0.176	0.174	0.166
	R	0.000	0.042	**0.188**	0.361	0.451	0.514	0.597	0.618	0.653	0.660
FreqWord _Bi	P	0.000	0.400	0.314	**0.303**	0.284	0.275	0.272	0.266	0.258	0.260
	R	0.000	0.056	0.076	**0.257**	0.347	0.382	0.410	0.438	0.444	0.479

Table 6.6 The classification performance with individual content features

Parameter		1	2	3	4	5	6	7	8	9	10
SigTerm _Uni&Bi	P	0.000	0.333	**0.348**	0.327	0.321	0.244	0.216	0.219	0.201	0.180
	R	0.000	0.042	**0.167**	0.250	0.347	0.458	0.583	0.604	0.604	0.639
CentroidVar _Uni&Bi	P	0.000	0.556	0.353	**0.250**	0.239	0.221	0.218	0.201	0.203	0.195
	R	0.000	0.035	0.083	**0.167**	0.236	0.299	0.375	0.403	0.528	0.632
FreqWord _Uni&Bi	P	0.000	0.381	**0.349**	0.306	0.296	0.255	0.205	0.180	0.172	0.176
	R	0.000	0.056	**0.153**	0.264	0.410	0.507	0.576	0.597	0.611	0.653
All_Con	P	0.000	0.316	**0.407**	0.345	0.276	0.253	0.239	0.219	0.205	0.197
	R	0.000	0.042	**0.167**	0.285	0.410	0.514	0.583	0.618	0.646	0.701

Table 6.7 The classification performance with combinational content features

6.3.3.3 Experiments on Event Features

We test our event-based summarization features under the learning-based framework. The best instance-based and concept-based event summarization approaches are employed to given weight scores of sentences, which are used as values of event features. The two event features are tested separately first, later on they are used together to given the final decision. The classification results are shown in Table 6.8. It can be seen that the performance of concept-based event summarization is better than the instance-based one generally.

Parameter		1	2	3	4	5	6	7	8	9	10
Event _Inst	P	0.000	0.000	0.346	**0.198**	0.187	0.181	0.160	0.122	0.114	0.112
	R	0.000	0.000	0.063	**0.181**	0.257	0.299	0.389	0.639	0.674	0.694
Event _Con	P	0.000	0.214	**0.436**	0.257	0.230	0.202	0.177	0.157	0.141	0.134
	R	0.000	0.021	**0.118**	0.326	0.417	0.493	0.535	0.618	0.667	0.701
All _Event	P	0.000	0.545	**0.344**	0.252	0.236	0.199	0.171	0.157	0.146	0.138
	R	0.000	0.042	**0.146**	0.243	0.354	0.465	0.590	0.639	0.674	0.694

Table 6.8 The classification performance with event features

6.3.3.4 Experiments on Relevance Features

To evaluate the role of relevance between sentences, we test the features come from links between sentences. Three kinds of relevance are exploited as three features and their performance is shown in Table 6.9. It can be seen that the most useful feature is DocFirstRel, i.e. how likely a sentence is similar to the first sentence of a document. On the other hand, the feature how likely a sentence is similar to the first sentence of a paragraph is not helpful. The performance of the combination of all relevance features is also the best. The precision is 0.488 and the recall is 0.146.

Parameter		1	2	3	4	5	6	7	8	9	10
DocFirstRel	P	0.000	**0.488**	0.488	0.488	0.488	0.169	0.159	0.150	0.140	0.130
	R	0.000	**0.146**	0.146	0.146	0.146	0.271	0.354	0.403	0.424	0.438
ParaFirstRel	P	0.000	0.000	0.000	0.000	0.000	0.000	0.000	**0.099**	0.086	0.086
	R	0.000	0.000	0.000	0.000	0.000	0.000	0.000	**0.319**	1.000	1.000
PageRankRel	P	0.000	0.000	0.000	0.000	**0.168**	0.151	0.145	0.127	0.118	0.114
	R	0.000	0.000	0.000	0.000	**0.208**	0.313	0.444	0.563	0.632	0.667
All_Rel	P	0.000	**0.488**	0.488	0.488	0.457	0.218	0.190	0.172	0.156	0.138
	R	0.000	**0.146**	0.146	0.146	0.146	0.326	0.451	0.556	0.604	0.646

Table 6.9 The classification performance with relevance features

6.3.4 Experiments on Combinational Feature Groups

After evaluating each group of features individually, we investigate their combinations. First every two feature groups are employed in the experiments, then every three groups and four groups of features. The performance on these feature groups is shown in Table 6.10, Table 6.11 and Table 6.12 respectively. The best performance on every two feature groups is achieved when surface features and event features are used. The precision and recall is 0.6 and 0.125. The performance comes from content features and relevance features is similar. It seems that all four types of features are useful. Surface, content and relevance features bring the best performance on every three feature groups. The precision is 0.595 and the recall is 0.174. When we use all four groups of features, the precision and recall is 0.579 and 0.153.

Parameter		1	2	3	4	5	6	7	8	9	10
Sur&Con	P	0.000	**0.575**	0.432	0.373	0.328	0.282	0.266	0.247	0.234	0.221
	R	0.000	**0.160**	0.264	0.347	0.458	0.514	0.590	0.639	0.674	0.694
Sur&Event	P	0.500	**0.600**	0.472	0.336	0.279	0.251	0.231	0.206	0.202	0.192
	R	0.007	**0.125**	0.236	0.313	0.396	0.514	0.597	0.618	0.694	0.736
Sur&Rel	P	0.000	**0.488**	0.450	0.320	0.299	0.256	0.256	0.248	0.234	0.226
	R	0.000	**0.146**	0.188	0.285	0.389	0.417	0.493	0.583	0.653	0.708
Con&Event	P	0.000	0.353	**0.384**	0.353	0.275	0.242	0.224	0.207	0.193	0.179
	R	0.000	0.042	**0.194**	0.340	0.424	0.521	0.583	0.625	0.653	0.660
Con&Rel	P	0.000	**0.588**	0.423	0.376	0.303	0.266	0.246	0.228	0.218	0.210
	R	0.000	**0.139**	0.229	0.326	0.424	0.493	0.563	0.604	0.646	0.701
Event&Rel	P	1.000	**0.543**	0.416	0.346	0.283	0.213	0.184	0.165	0.153	0.145
	R	0.007	**0.132**	0.222	0.257	0.361	0.535	0.604	0.618	0.632	0.653

Table 6.10 The classification performance with two feature groups

Parameter		1	2	3	4	5	6	7	8	9	10
Sur&Con &Event	P	1.000	**0.595**	0.434	0.375	0.321	0.276	0.253	0.239	0.221	0.221
	R	0.007	**0.153**	0.250	0.354	0.472	0.514	0.590	0.646	0.674	0.715
Sur&Con &Rel	P	0.000	**0.595**	0.434	0.389	0.330	0.282	0.261	0.249	0.235	0.228
	R	0.000	**0.174**	0.250	0.354	0.465	0.521	0.583	0.632	0.667	0.688
Sur&Event &Rel	P	1.000	**0.553**	0.436	0.312	0.282	0.255	0.229	0.222	0.215	0.197
	R	0.007	**0.146**	0.236	0.299	0.403	0.493	0.542	0.625	0.694	0.722
Con&Event &Rel	P	0.000	**0.581**	0.385	0.376	0.289	0.257	0.221	0.217	0.198	0.192
	R	0.000	**0.125**	0.208	0.347	0.424	0.507	0.549	0.604	0.632	0.660

Table 6.11 The classification performance with three feature groups

Parameter		1	2	3	4	5	6	7	8	9	10
Sur&Con& Event&Rel	P	1.000	**0.579**	0.424	0.372	0.316	0.279	0.251	0.243	0.229	0.218
	R	0.007	**0.153**	0.250	0.354	0.465	0.528	0.590	0.639	0.667	0.694

Table 6.12 The classification performance with all feature groups

6.3.5 Experiments on ROUGE Evaluations

ROUGE is employed to evaluate groups of features and the corresponding results are shown in Table 6.13. From this table, we can see surface features and relevance features

achieve best ROUGE-1 performance, and then content features and event features. The performance of event features is comparable to that of content features. Please note that there are six content features and just two event features. The best results in all experiments (ROUGE-1 is 0.396) are close to the upper bound 0.422, which is achieved from manual summaries. It can be seen that the learning-based framework and employed features are successive for extractive summarization.

	Precision	Recall	ROUGE-1	ROUGE-2	ROUGE-L
Sur	0.488	0.146	0.373	0.103	0.356
Con	0.407	0.167	0.352	0.074	0.334
Event	0.344	0.146	0.344	0.064	0.325
Rel	0.488	0.146	0.373	0.103	0.356
Sur+Con	0.575	0.160	0.380	0.109	0.363
Sur+Event	0.600	0.125	0.348	0.091	0.332
Sur+Rel	0.488	0.146	0.373	0.103	0.356
Con+Event	0.384	0.194	0.344	0.071	0.330
Con+Rel	0.588	0.139	0.375	0.103	0.358
Sur+Con+Event	0.595	0.153	0.379	0.106	0.363
Sur+Con+Rel	**0.595**	**0.174**	**0.396**	**0.116**	**0.374**
Con+Event+Rel	0.581	0.125	0.371	0.101	0.353
Sur+Con+Event+Rel	0.579	0.153	0.375	0.106	0.359

Table 6.13 The ROUGE results on each feature group

6.3.6 Experiments on Re-ranking

When the total number of the words in the important sentences labeled by a SVM classifier is far beyond or far away from 200 words, i.e. the length limitation of summaries required, the re-ranking procedure becomes necessary. To investigate which feature or features are suitable as re-ranking criteria, we design six different schemes based on different surface features and content features. The results are given in Table 6.14. It can be seen that the best performance is achieved when length and position of sentences are used together as the re-ranking criteria.

	ROUGE-1	ROUGE-2	ROUGE-L
Length	0.383	0.111	0.352
Position	0.383	0.108	0.346
Length+Position	**0.394**	**0.116**	**0.357**
FreqWord_Uni	0.382	0.111	0.341
FreqWord_Bi	0.381	0.110	0.341
FreqWord_Uni&Bi	0.384	0.114	0.343

Table 6.14 The ROUGE results from different re-ranking schemes

6.4 Discussion

The centroid features presented in Section 6.2.2 are not exactly same as discussed in [Radev et al. 2004]. We compute the *idf* factor over the documents within a cluster in stead of over all clusters. The precision and recall of the latter are 0.299 and 0.139. Corresponding ROUGE-1 and ROUGE-2 scores are 0.361, and 0.085 respectively. ROUGE results of centroid features used in this way are comparable with other content features. However, the results become worse when they are combined with other features.

In our preliminary experiments with signature word features, better ROUGE results are achieved if we count number of them, rather than sum weights of them. When we further explore the other two content features, we find it also applies to the centroid features although not to high frequency word features (see Table 6.15). However, when we combine the content features calculated by numbers with other kinds of features, no improvement is achieved at all. This is a strange result to us and need more efforts to study in the future.

	ROUGE-1	ROUGE-2	ROUGE-L
Centroid_Uni&Bi	0.361	0.085	0.325
Centroid_Uni&Bi_Num	0.379	0.086	0.345
SigTerm_Uni&Bi	0.356	0.059	0.318
SigTerm_Uni&Bi_Num	0.361	0.075	0.329
FreqWord_Uni&Bi	0.366	0.081	0.327
FreqWord_Uni&Bi_Num	0.350	0.074	0.318

Table 6.15 The ROUGE results on unigram and bigram content features

To compare with the learning based approach, we also conduct linear combination based summarization approach. All summarization features are used and they are same as those involved in learning based approach. The test data and evaluation tool are same also. We set the weight parameters for each feature as equal experimentally. The Rouge-1 and Rouge-2 scores are 0.362 and 0.008 respectively for linear combination based approach. However, they are 0.375 and 0.106 for leaning based approach (Table 6.13). This comparison partially shows that learning based summarization approach is better than linear combination based one.

6.5 Chapter Summary

We employ a learning-based framework to extract important sentences by exploiting surface, content, relevance and event features. As we discussed in Section 6.2, binary classification is chose to fulfill this task. Different re-ranking schemes are proposed and the best one is based on two surface features. Under this framework, experiments show that our feature combinations achieved competitive result. The highest performance in our experiments is achieved by the combination of surface, content and relevance features. The ROUGE-1 score is 0.396. It is a competitive result on this data set.

Experiments show that relevance features or surface features are able to achieve highest performance individually (ROUGE-1 0.373). Based on the combination of surface features and content features, if relevance features are incorporated the ROUGE-1 score can be improved obviously from 0.380 to 0.396. It can be seen that relevance features are important to extractive summarization. We also find that the performance from event features is comparable to that of content features.

Chapter 7

Conclusion

The major objective of this study is to investigate summarization techniques with the help of temporal information. There are three main problems to be resolved. The first one is how to represent contents of documents. The second one is how to identify important contents in documents. Temporal information is a possible source to be exploited in this procedure. The third one is how to identify and normalize temporal expressions to get temporal information.

Event is natural unit to represent meanings embedded in documents. Extractive summarization approaches based on event instance and event concept are investigated. In this study event is defined as event terms and associated event elements at sentence level. Event terms are verbs and action nouns, while event elements contain four types of named entities, such as person names, organization names, locations and times. One occurrence of an event term/element in a document is an instance of the event term/element, while the collection of the same event term/element is a concept of the event term/element. Two kinds of summarization approaches are proposed based on event instance and concept respectively.

Independent extractive summarization approaches based on event instance are studied first, and then relevant one. Documents often narrate more than one similar or related event. Therefore it can be seen that the relevance between events is a possible source to be exploited to improve the summarization performance. In the independent approach the contents of documents are represented by event instances and an event instance is weighted by its frequency. The weight of each sentence is the sum of weights of event instances. In relevant approach the contents of documents are also represented by event instances, but

same instances are connected and an event map is built. A suitable link analysis algorithm – PageRank is used to evaluate the weight of each node in this map, i.e. each event instance. Finally each sentence is weighted according to the event instances contained in them.

ROUGE is employed in experiments of this study to measure unigram or bigram overlap between machine-generated summaries and model summaries. The ROUGE-1 score of independent and relevant instance-based event summarization approaches are 0.315 and 0.334 respectively. It can be seen that the relevance is useful to improve the quality of summaries. Therefore just relevant concept-based event summarization approaches are investigated.

The relevance between two concepts of event term/element may be intra-event or inter-event. Here intra-event relevance means two concepts occur in same event. One kind of inter-event relevance comes from the similarity of two concepts in WordNet. Another kind of inter-event relevance exists between every two event terms/elements which co-occur with the other event term/element. Other inter-event relevance is also investigated, which are based on lexical similarity, clustering and window size. Different relevance schemes are compared. The best performance is from the combination of the intra-event relevance and the inter-event relevance which lies between event elements co-occurred with same event terms. The ROUGE-1 score is 0.352. The relevant concept-based approach achieves better results, compared with the relevant instance-based approach.

As the limitation of natural language processing techniques nowadays, general summarization approaches suitable to all types of document are nearly impossible. It is obvious that topics may shift over time in news documents. For example, the topic of a cluster of documents in DUC 2001 data collection is fires in USA. Different descriptions about fires at different years are given. This characteristic can be exploited to improve summarization approaches. The performance of event-based summarization approaches is promising in our previous study. Therefore the function of temporal information is explored based on event summarization. Event concept is a set of instances which occur on different

times, so it is not suitable to be improved by temporal information. Therefore independent instance-based summarization is selected as the baseline.

To filter errors introduced by temporal information processing, this information is tagged manually. Event instances are anchored on certain days of the time line and their importance is evaluated from local and global points of views. It is assumed that important events are those occurring frequently in certain periods and mainly within certain periods. Two statistical measures, temporal *tf*idf* and χ^2, are employed. Each of them can be used to evaluate the association of event instances to a certain period and width of event instance distribution on the time line. The importance of each sentence is evaluated by the sum of the weights of all the event instances contained in it. Then sequential and round robin sentence selection strategies are compared. The combination of *tf*idf* and sequential sentence selection by sentence weight performs best. Compared with event-based summarization without considering temporal distribution, it improves ROUGE-1 by 18.8% in the preliminary experiments. This approach also achieves significant improvement when evaluated by a subject. The experiment on documents with auto-tagged temporal information also shows the improvement.

It's difficult to extract sentences according to one universal feature of sentences. Kinds of features are proposed in related work and this study. To combine these features effectively a learning-based frame work is designed. Event-based summarization approaches is employed to generate event features of sentences. Surface, content and relevance features are also used to decide the importance of sentences. Surface features reflect external characteristics of sentences, such as their position in documents. Content features refer to the use of frequent or significant words. Relevance features refer to those less easily observed features of a sentence, such as the relevance between a sentence and the first sentence of a document. Experiments show that the combination of surface, content and relevance features achieves the best performance. The ROUGE-1 score of these features is 0.396. It also can be

seen that the performance from event features is comparable to that of content features. It seems that event-based approaches are applicable ways for summarization.

Temporal information processing is valuable in many NLP applications, for example, document summarization. Contents can be anchored on the time line and their distributions can be exploited to identify the focus. Temporal expressions convey crucial information for the anchoring. Temporal expressions are defined as chunks of text which convey knowledge about time point or duration. According to TIMEX2 guidelines, temporal expressions include dates, times of day, durations, frequency expressions, event-anchored expressions, and so on. To retrieve useful temporal information, the extents of temporal expressions in raw text should be identified and then temporal attributes of the expressions should be explained according to the guideline. An English system, which is rule-based with the help of constraints, is designed for the two tasks. The evaluation results of this system are comparable to those of English systems at TERN 2004. A Chinese system is also designed and its evaluation results are the best among those of Chinese systems at TERN 2004.

In short, this study has made five major contributions:

1. Instance-based event summarization approaches are proposed to identify the focus of a cluster of documents. The context of event term, such as the number and type of named entities, is investigated in independent approaches. It is found that the context does not influence the summarization performance heavily. It is also found that independent event-based summarization approaches are suitable to generate longer summaries. Relevant instance-based approaches build an event map and then use PageRank to evaluate the importance of event instances. Compared with independent approaches, it achieves significant improvement.
2. An event concept is a set of event instances. Relevant concept-based event summarization approaches are proposed to investigate different types of relevance between event concepts. Intra-event and inter-event relevance are considered.

Experiments show that the inter-event relevance between event element and event element achieves the best results with the help of intra-event relevance.

3. To identify the important contents of documents which contain topics shifting over time, temporal–oriented event-based summarization is investigated. Event instances are anchored on the time line. Two statistics, *tf***idf* and χ^2, are used to evaluate the importance of instances. Round robin and sequential sentence selection strategies are proposed to generate the final summary. Compared with event-based summarization without considering time, *tf***idf* and sequential selection strategy achieve significant improvement.
4. A universal learning-based framework is proposed to incorporate multiple sentence features. Surface, content, event and relevance features are investigated. Experiments show that surface and relevance features achieve best performance when just one type of features is used. The performance of event features is comparable to that of content features. The best results of this framework (ROUGE-1 is 0.396) are close to the upper bound 0.422, which is achieved from manual summaries.
5. Temporal expression extraction and normalization systems are designed for English and Chinese. These two systems are based on grammar rules and constraint rules. They are easily to be adapted to other languages. The Chinese system achieves the highest performance at TERN 2004 and the performance of English system is in the middle of all evaluation results.

Bibliography

ACE-2005. 2005. http://www.nist.gov/speech/tests/ace/ace05/

Afantenos, S.D., Karkaletsis V. and Stamatopoulos P. 2005. Summarizing reports on evolving events; Part I: linear evolution. In Proceedings of Recent Advances in Natural Language Processing (RANLP-2005), pages 18-24, Borovets, Bulgaria.

Ahn, D., Adafre, S. F. and Rijke, M. 2005. Towards task-based temporal extraction and recognition. In Proceedings of Dagstuhl Workshop on Annotating, Extracting, and Reasoning about Time and Events, Dagstuhl, Germany.

Allan, J., Gupta, R. and Khandelwal, V. 2001. Temporal summaries of news topics. In Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR-2001), pages 10-18, New Orleans, USA.

Allen, J.F. 1981. An interval-based representation of temporal knowledge. In Proceedings of the 7th International Joint Conference on Artificial Intelligence (IJCAI-1981), pages 221-226, Vancouver, Canada.

Allen, J. F. 1984. Towards a general theory of action and time. Artificial Intelligence, vol. 23, no. 2, pages 123-154.

Aone, C., Okurowski, M.E., Gorlinsky, J. and Larsen, B. 1999. A trainable summarizer with knowledge acquired from robust NLP techniques. Advances in Automatic Text Summarization. MIT Press, Cambridge, Massachusetts, pages 71-80.

Barzilay, R. and Elhadad, M. 1997. Using lexical chains for text summarization. In Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and the 8th European Chapter of the Association for Computational Linguistics (ACL/EACL-1997) Workshop on Intelligent Scalable Text Summarization, pages 10-17, Madrid, Spain.

Barzilay, R. and Elhadad, M. 1999a. Using lexical chains for text summarization. Advances in Automatic Text Summarization. MIT Press, Cambridge, Massachusetts, pages 111-121.

Barzilay, R., Elhadad, M. and Mckeown, K. 1999b. Information fusion in the context of multi-document summarization. In Proceedings of 37th Annual Meeting of the Association for Computational Linguistics (ACL-1999), pages 550-557, College Park, Maryland, USA.

Barzilay, R., Elhadad, M. and Mckeown, K. 2001. Sentence ordering in multi-document summarization. In Proceedings of the 1st International Conference on Human Language Technology (HLT-2001), pages 32-38, San Diego, USA.

Barzilay, R. and Lapata, M. 2005. Modeling local coherence: an entity-based approach. In Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL-2005), pages 141-148, Ann Arbor, USA.

Baxendale, P.B. 1958. Man-made index for technical literature -- an experiment. IBM Journal of Research and Development, vol 2, no. 4, pages 354-361.

Boguraev, B. and Kennedy, C. 1997. Salience-based content characterization of text documents. In Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and the 8th European Chapter of the Association for Computational Linguistics (ACL/EACL-1997) Workshop on Intelligent Scalable Text Summarization, pages 2-9, Madrid, Spain.

Brandow, R., Mitze, K. and Rau, L. 1995. Automatic condensation of electronic publications by sentence selection. Information Processing and Management, vol. 31, no. 5, pages 675-686.

Buckley, C. and Cardie, C. 1997. Using EMPIRE and SMART for high-precision IR and summarization. In Proceedings of the TIPSTER Text Phase III 12-month Workshop, San Diego, USA.

Carbonell, J., Y. Geng and Goldstein, J. 1997. Automated query-relevant summarization and diversity-based reranking. In Proceedings of the 15th International Joint Conference on Artificial Intelligence Workshop on AI in Digital Libraries (IJCAI-1997), pages 12-19, Nagoya, Japan.

Carbonell, J. and Goldstein, J. 1998. The use of MMR, diversity-based reraning for reordering documents and producing summaries. In Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retreval (SIGIR-1998), pages 335-336, Melbourne, Australia.

Chang C.-C. and Lin C.-J. 2001. LIBSVM: a library for support vector machines. http://www.csie.ntu.edu.tw/~cjlin/libsvm

Conroy, J. and O'Leary D. 2001. Text summarization via Hidden Markov Models. In Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR-2001), pages 406-407, New Orleans, USA.

Conroy, J. and Schlesinger, J. 2004. Left-brain/right-brain multi-document summarization. http://duc.nist.gov/pubs.html

Cunningham, H., Maynard, D., Bontcheva, K. Tablan, V. 2002. GATE: a framework and graphical development environment for robust NLP tools and applications. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL-2002), pages 168-175, Philadelphia, USA.

Daniel, N., Radev, D.R. and Allison, T. 2003. Sub-event based multi-document summarization. In Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT/NAACL-2003) Workshop on Text Summarization, pages 9-16, Edmonton, Canada.

DeJong, G.F. 1978. Fast skimming of news stories: the FRUMP system. Ph.D. thesis, Yale University, New Haven, USA.

DeJong, G. F. 1979. Skimming stories in real time: an experiment in integrated understanding (Department of Computer Science Research Report #156). Yale University, New Haven, USA.

Dunning, T. 1993. Accurate methods for the statistics of surprise and coincidence. Computational Linguistics. vol. 19, no. 1, pages 61-74.

Edmundson, H.P. 1969. New methods in automatic extracting. Journal of the Association for Computing Machinery, vol. 16, no. 2, pages 264-285.

Erkan, G. and Radev, D.R. 2004. LexRank: graph-based lexical centrality as salience in text summarization. Journal of Artificial Intelligence Research, vol. 22, pages 457-479.

Estela, S., Martinez-Barco, P. and Munoz, R. 2002. Recognizing and tagging temporal expressions in Spanish. In Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC-2002) Workshop on Annotation Standards for Temporal Information in Natural Language, pages 44-51, Las Palmas, Spain.

Fellbaum, C. 1998. WordNet: An Electronic Lexical Database. MIT Press.

Ferro, L., Gerber, L., Mani, I., Sundheim, B. and Wilson, G. 2004. TIDES 2003 standard for the annotation of temporal expressions. http://timex2.mitre.org

Filatova E. and Hovy E. 2001. Assigning time-stamps to event-clauses. In Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics (ACL-2001) Workshop on Temporal and Spatial Information Processing, pages 88-95, Toulouse, France.

Filatova, E. and Hatzivassiloglou, V. 2004. Event-based extractive summarization. In Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-2004) Workshop on Summarization, pages 104-111, Barcelona, Spain.

Gerber, L., Huang, S. and Wang, X. 2004. TIDES 2003 standard for the annotation of temporal expressions (Chinese supplement draft). http://timex2.mitre.org

Hahn, U. and Reimer, U. 1999. Knowledge-based text summarization: salience and generalization operators for knowledge base abstraction. Advances in Automatic Text Summarization. MIT Press, Cambridge, Massachusetts, pages 215-232.

Hartley, J., Sydes, M. and Blurton, A. 1996. Obtaining information accurately and quickly: are structured abstracts more efficient? Journal of Information Science, vol. 22, no. 5, pages 349-356.

Hearst, M. 1994. Multi-paragraph segmentation of expository text. In Proceedings of the 32th Annual Meeting of the Association for Computational Linguistics (ACL-1994), pages 9-16, Las Cruces, USA.

Hobbs, J. R. and Pan, F. 2004. An ontology of time for the semantic web. ACM Transactions on Asian Language Information Processing, vol. 3, no. 1, pages 66-85.

Jatowt, A. and Ishizuka, M. 2004. Temporal web page summarization. In Proceedings of the 5th International Conference on Web Information Systems Engineering (WISE-2004), pages 303-312, Brisbane, Australia.

Jahna, O., Radev, D.R. and Luo, A. 2002. Revisions that improve cohesion in multi-document summaries: a preliminary study. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL-2002) Workshop on Automatic Summarization, pages 27-36, Philadelphia, USA.

Jang, S.B., Baldwin, J. and Mani, I. 2004. Automatic TIMEX2 tagging of Korean news. ACM Transactions on Asian Language Information processing, vol. 3, no. 1, pages 51-65.

Jing, H. and McKeown, K. 1999. The decomposition of human-written summary sentences. In Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR-1999), pages 129-136, Berkeley, USA.

Knight, K and Marcu, D. 2000. Statistics-based summarization--step one: sentence compression. In Proceedings of the 17th National Conference of the American Association for Artificial Intelligence (AAAI-2000), pages 703-710, Austin, USA.

Kupiec, J., Pedersen, J. O. and Chen, F. 1995. A trainable document summarizer. In Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 68-73, Seattle, United States.

Lehnert, W. G. 1982. Plot units: a narrative summarization strategy. Strategies for Natural Language Processing, Lehnert, W. G. and Ringle, M.H. (Eds.). Lawrence Erlbaum Associates, Hillsdale, New Jersey, USA.

Leskovec, J., Grobelnik, M. and Milic-Frayling N. 2004. Learning sub-structures of document semantic graphs for document summarization. In Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (SIGKDD-2004) Workshop on Link Analysis and Group Detection, Seattle, USA.

Li, B., Li, W., Lu, Q. and Wu, M. 2005. Profile-based event tracking. In Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR-2005), pages 631-632, Salvador, Brazil.

Li, W., Wong, K.-F. and Yuan, C. 2001. A model for processing temporal references in Chinese. In Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics (ACL-2001) Workshop on Temporal and Spatial Information Processing, pages 33-40, Toulouse, France.

Li, W., Wu, M., Lu, Q. and Wong K.-F. 2006a. Integrating temporal distribution information into event-based summarization. International Journal of Computer Processing of Oriental Languages, vol. 19, no. 2-3, pages 201-222.

Li, W., Xu, W., Wu, M., Yuan, C. and Lu, Q. 2006b. Extractive summarization using inter- and intra- event relevance. In Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (COLING/ACL-2006), pages 369-376, Sydney, Australia.

Lim, J.-M., Kang, I.-S., Bae J.-H., Lee J.-H. 2004. Sentence extraction using time features in multi-document summarization. In Proceedings of the 1st Asia Information Retrieval Symposium (AIRS-2004), pages 82-93, Beijing, China.

Lin, C.Y. 1999. Training a selection function for extraction. In Proceedings of the 8th Annual International ACM Conference on Information and Knowledge Management (CIKM-1999), pages 55-62, Kansas City, USA.

Lin, C.Y. and Hovy, E. 1997. Identifying topics by position. In Proceedings of 5th Conference on Applied Natural Language Processing (ANLP-1997), pages 283-290, Washington, D.C., USA.

Lin, C.Y. and Hovy, E. 2000. The automated acquisition of topic signatures for text summarization. In Proceedings of the 18th International Conference on Computational Linguistics (COLING-2000), pages 495-501, Saarbrücken, Germany.

Lin, C.Y. and Hovy, E. 2002. From single to multi-document summarization: a prototype system and its evaluation. In Proceedings of the 40th Conference of the Association of Computational Linguistics, pages 457-464, Philadelphia, USA.

Lin, C.Y. and Hovy, E. 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics. In Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT/NAACL-2003), pages 71-78, Edmonton, Canada.

Liu, M., Li, W., Wu, M. and Hu, H. 2007a. Event-based extractive summarization using event semantic relevance from external linguistic resource. In Proceedings of the 6th International Conference on Advanced Language Processing and Web Information Technology (ALPIT-2007), Luoyang, China.

Liu, M. Li, W., Wu M. and Lu, Q. 2007b. Extractive summarization based on event term clustering. In Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL-2007), Prague, Czech Republic, pages 185-188.

Luhn, H. 1958. The automatic creation of literature abstracts. IBM Journal of Research and Development, vol. 2, no. 2, pages 159-165.

Mani, I. 2001. Automatic Summarization. John Benjamin's Publishing Company, Amsterdam.

Mani, I. and Bloedorn, E. 1997. Multi-document summarization by graph search and matching. In Proceedings of the 14th National Conference on Artificial Intelligence (AAAI-1997), pages 622-628, Providence, USA.

Mani, I. and Bloedorn, E. 1999. Summarizing similarities and differences among related documents, Information Retrieval, vol. 1, no. 1-2, pages 35-67.

Mani, I., Gates, B. and Bloedorn, E. 1999. Improving summaries by revising them. In Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL-1999), pages 558-565, College Park, Maryland, USA.

Mani, I. and Maybury, M. T. (eds.). 1999. Advances in automatic text summarization. MIT Press, Cambridge, Massachusetts.

Mani, I., Pustejovsky, J. and Sundheim, B. 2004. Introduction to the special issue on temporal information processing. ACM Transactions on Asian Language Information Processing, vol. 3, no. 1, pages 1-10.

Mani, I. and Wilson G. 2000. Robust temporal processing of news. In Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (ACL-2000), pages 69-76, New Brunswick, New Jersey, USA.

Mann, W. and Thompson, S. 1988. Rhetorical structure theory: toward a functional theory of text organization. Text, vol. 8, no. 3, pages 243-281.

Marcu, D. 1997a. The rhetorical parsing of natural language texts. In Proceedings of the 35th Annual Meeting of the Association for computational Linguistics (ACL-1997), pages 96-103, Madrid, Spain.

Marcu, D. 1997b. The Rhetorical parsing, summarization, and generation of natural language texts. Ph.D. thesis. Department of Computer Science, University of Toronto, Canada.

Marcu, D. 1999. Discourse trees are good indicators of importance in text. Advances in Automatic Text Summarization. MIT Press, Cambridge, Massachusetts, pages 123-136.

Marcu, D. and Gerber, L. 2001. An inquiry into the nature of multi-document abstracts, extracts, and their evaluation. In Proceedings of the 2nd Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL-2001) Workshop on Automatic Summarization, pages 1-8, Pittsburgh, USA.

Maybury, M.T. 1991. Planning multi-sentential English text using communicative acts. Ph.D. thesis, University of Cambridge, England.

Maybury, M.T. 1992. Communicative acts for explanation generation. International Journal for Nam-Machine Studies, vol. 37, no. 2, pages 135-172.

Maybury, M.T. 1995. Generating summaries from event data. International Journal of Information Processing and Management: Special Issue on Text Summarization, vol. 31, no. 5, pages 735-751.

McKeown, K., Robin, J. and Kukich, K. 1995. Generating concise natural language summaries. Journal of Information Processing and Management, vol 31, no. 5, pages 703-733.

Myaeng S.H. and Jang D. 1999. Development and evaluation of a statistically based document summarization system. Advances in Automatic Text Summarization. MIT Press, Cambridge, Massachusetts, pages 61-70.

Mihalcea, R. 2005. Language independent extractive summarization. In Proceedings of the 43rd Annual Meeting of the Association for computational Linguistics (ACL-2005) on Interactive Poster and Demonstration Sessions, pages 49-52, Ann Arbor, USA.

MUC-7. 1998. http://www-nlpir.nist.gov/related_projects/muc/.

Nenkova, A., Vanderwende, L. and McKeown, K. 2006. A compositional context sensitive multi-document summarizer: exploring the factors that influence summarization. In Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR-2006), pages 573-580, Seattle, USA.

Ono, K., Sumita, K. and Miike, S. 1994. Abstract generation based on rhetorical structure extraction. In Proceedings of the 15th International Conference on Computational Linguistics (COLING-1994), pages 344-348, Kyoto, Japan.

Page, L, Brin, S., Motwani, R. and Winograd, T. 1998. The PageRank Citation Ranking: Bring Order to the Web. Technical Report, Stanford University, USA.

Pedersen, T., Patwardhan, S. and Michelizzi, J. 2004. WordNet :: similarity – measuring the relatedness of concepts. In Proceedings of the 19th National Conference of the American Association for Artificial Intelligence (AAAI-2004), pages 25-29, San Jose, USA.

Pollock, J. J. and A. Zamora. 1975. Automatic abstracting research at chemical abstracts service. Journal of Chemical Information and Computer Sciences, vol. 15, no. 4, pages 226-232.

Radev, D.R., Allison, T., Blair-Goldensohn, S., Blitzer, J., Celebi, A., Dimitrov, S., Drabek, E., Hakim, A., Lam, W., Liu, D., Otterbacher, J., Qi, H., Saggion, H., Teufel, S., Topper, M., Winkel, A. and Zhang, Z. 2004. MEAD - a platform for multi-document multilingual text summarization. In Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC-2004), Lisbon, Portugal.

Radev, D.R., Hovy, E. and McKeown, K. 2002. Introduction to special issue on summarization. Computational Linguistics, vol 28, no.4, pages 399-408.

Radev, D.R., Jing, H. and Budzikowska, M. 2000a. Summarization of multiple documents: clustering, sentence extraction, and evaluation. In Proceedings of the 6th Applied Natural Language Processing Conference and 1st Meeting of the North American Chapter of the Association for Computational Linguistics (ANLP/NAACL-2000) Workshop on Automatic Summarization, Seattle, USA.

Radev, D.R., Jing, H. and Budzikowska, M. 2000b. Centroid-based summarization of multiple documents: sentence extraction, utility-based evaluation, and user studies. In Proceedings of the 6th Applied Natural Language Processing Conference and 1st

Meeting of the North American Chapter of the Association for Computational Linguistics (ANLP/NAACL-2000) Workshop on Summarization, pages 21-29, Seattle, USA.

Rino, L. and Scott, D. 1994. Content selection in summary generation. In Proceedings of the 3th International Conference on the Cognitive Science of Natural Language Processing, Dublin, Ireland.

Rau, L. and Jacobs, P. 1991. Creating segmented databases from free text for text retrieval. In Proceedings of the 14th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, pages 337-346, Chicago, USA.

Rennie, D. and Glass, R.M. 1991. Structuring abstracts to make them more informative. Journal of the American medical Association, vol. 266, no. 1, pages 116-117.

Rich, E. and Knight, K. 1991. Artificial Intelligence (Second Edition). McGraw-Hill, New York.

Salton, G. and McGill, M.J. 1983. Introduction to Modern Information Retrieval. McGraw-Hill, New York.

Salton, G., Singhal, A., Mitra, M. and Buckley, C. 1997. Automatic text structuring and summarization. Information Processing and Management, vol. 33, no. 2, pages 193-207.

Sauri, R., Littman, J., Knippen, B., Gaizauskas, R., Setzer, A. and Pustejovsky, J. 2004. TimeML Annotation Guidelines. http://cs.brandeis.edu

Schilder, F. and Habel, C. 2001. From temporal expressions to temporal information: semantic tagging of news messages. In Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics (ACL 2001) Workshop on Temporal and Spatial Information Processing, pages 65-72, Toulouse, France.

Sparck-Jones, K. 1999. Automatic summarizing: factors and directions. Advances in Automatic Text Summarization. MIT Press, Cambridge, Massachusetts, pages 1-12.

Strzalkowski, T., Stein, G., Wang, J. and Wise, B. 1999. A robust practical text summarizer. Advances in Automatic Text Summarization. MIT Press, Cambridge, Massachusetts, pages 137-154.

Swan, R. and Allan, J. 2000. Automatic generation of overview timelines. In Proceedings of the 23th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR-2000), pages 49-56, Athens, Greece.

TDT-1998. 1998. http://www.nist.gov/speech/tests/tdt/

TERN-2004. 2004. http://timex2.mitre.org/tern.html

Teufel, S. and Moens, M. 1999. Argumentative classification of extracted sentences as a first step towards flexible abstracting. Advances in Automatic Text Summarization. MIT Press, Cambridge, Massachusetts, pages 137-154.

Teufel, S. and Moens, M. 2002. Summarizing scientific articles: experiments with relevance and rhetorical status. Computational Linguistics, vol. 28, no. 4, pages 409-445.

Vanderwende, L., Banko, M. and Menezes, A. 2004. Event-centric summary generation. Available at http://duc.nist.gov/pubs.html

Vazov, N. 2001. A system for extraction of temporal expressions from French texts based on syntactic and semantic constraints. In Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics (ACL-2001) Workshop on Temporal and Spatial Information Processing, pages 96-103, Toulouse, France.

Witbrock, M. and Mittal, V. 1999. Ultra-summarization: a statistical approach to generating highly condensed non-extractive summaries. In Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR-1999), pages 315-316, Berkeley, USA.

Wu, M. 2006. Investigation on event-based summarization. In Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (COLING/ACL-2006) Student Research Workshop, pages 37-42, Sydney, Australia.

Wu, M., Li, W., Chen, Q. and Lu, Q. 2005a. Normalizing Chinese temporal expressions with multi-label classification. In Proceedings of 2005 IEEE International Conference on Natural language Processing and Knowledge Engineering (NLPKE-2005), pages 318 - 323, Wu Han, China.

Wu, M., Li, W., Lu, Q. and Li, B. 2005b. CTEMP: A Chinese temporal parser for extracting and normalizing temporal information. In Proceedings of the 2nd International Joint Conference on Natural Language Processing (IJCNLP-2005), pages 694-706, Jeju Island, South Korea.

Wu, M., Li, W., Lu, Q. and Wong, K.-F. 2006. Chinese temporal expression extraction and normalization. Submitted to ACM Transactions on Asian Language Information Processing.

Wu, M., Li, W., Lu, Q. and Wong, K.-F. 2007a. Event-based summarization using time features. In Proceedings of the 8th International Conference on Intelligent Text Processing and Computational Linguistics (CICLING-2007), pages 563-574, Mexico City, Mexico.

Wu, M., Li, W., Lu, Q. and Wong, K.-F. 2007b. Exploiting surface, content and discourse features for extractive summarization. To appear in the Proceedings of 2007 IEEE International Conference on Natural Language Processing and Knowledge Engineering (NLPKE-2007), Beijing, China.

Xu, W., Li, W., Wu, M., Li, W. and Yuan, C. 2006. Deriving event relevance from the ontology constructed with formal concept analysis. In Proceedings of the 7th International Conference on Intelligent Text Processing and Computational Linguistics (CICLING-2006), pages 480-489, Mexico City, Mexico.

Yoshioka, M. and Haraguchi, M. 2004. Multiple news articles summarization based on event reference information. In Working Notes of the 4th National Institute of Informatics Test Collection for IR Systems (NTCIR-2004) Workshop Meeting, Tokyo, Japan.