MULTI-DOCUMENT SUMMARIZATION BASED ON ATOMIC SEMANTIC EVENTS AND THEIR TEMPORAL RELATIONS

MD MOHSIN UDDIN
Bachelor of Science in Computer Science and Engineering
Bangladesh University of Engineering and Technology (Bangladesh), 2011

A Thesis
Submitted to the School of Graduate Studies
of the University of Lethbridge
in Partial Fulfillment of the
Requirements for the Degree

MASTER OF SCIENCE

Department of Mathematics and Computer Science
University of Lethbridge
LETHBRIDGE, ALBERTA, CANADA

MULTI-DOCUMENT SUMMARIZATION BASED ON ATOMIC SEMANTIC EVENTS AND THEIR TEMPORAL RELATIONS

MD MOHSIN UDDIN

Approved:

Name	*Signature*	*Rank*	*Date*
Dr. Yllias Chali Supervisor:			
Dr. Wendy Osborn Committee Member:			
Dr. John Zhang Committee Member:			
Dr. Howard Cheng Chair, Thesis Examination Committee:			

This thesis is dedicated to my beloved mother whose support, inspiration, and constant love have sustained me throughout my life.

Abstract

Automatic multi-document summarization (MDS) is the process of extracting the most important information such as events and entities from multiple natural language texts focused on the same topic. We extract all types of semantic atomic information and feed them to a topic model to experiment with their effects on a summary. We design a coherent summarization system by taking into account the sentence relative positions in the original text. Our generic MDS system has outperformed the best recent multi-document summarization system in DUC 2004 in terms of ROUGE-1 recall and f_1-measure. Our query-focused summarization system achieves a statistically similar result to the state-of-the-art unsupervised system for DUC 2007 query-focused MDS task in ROUGE-2 recall measure. Update Summarization is a new form of MDS where novel yet salience sentences are chosen as summary sentences based on the assumption that the user has already read a given set of documents. In this thesis, we present an event based update summarization where the novelty is detected based on the temporal ordering of events and the saliency is ensured by event and entity distribution. To our knowledge, no other study has deeply investigated the effects of the novelty information acquired from the temporal ordering of events (assuming that a sentence contains one or more events) in the domain of update MDS. Our update MDS system has outperformed the state-of-the-art update MDS system in terms of ROUGE-2, and ROUGE-SU4 recall measures. Our MDS systems also generate quality summaries which are manually evaluated based on popular evaluation criteria.

Acknowledgments

All praise be to the Almighty who gave me the strength to overcome all obstacles in my journey of life. I would like to thank, first of all, my supervisor Professor Dr. Yllias Chali for more than a few reasons. He not only introduced me to this wonderful and fascinating world of Natural Language Processing, but also provided me with numerous resources and helpful directions. Without his constant supervision and words of encouragement, this thesis would not have been possible at all. I also want to thank the members of my thesis committee: Dr. Wendy Osborn, and Dr. John Zhang for their valuable suggestions. I would like to express my deep gratitude to all the members of our research group for their valuable suggestions and continual encouragements. My special thanks goes to Dr. Sadid Hasan and Cody Rioux, my colleagues in the research group. My mother and all of my relatives supported me and encouraged me to the best of their ability. My heart-felt gratitude goes to them.

Contents

List of Tables

Chapter 1

Introduction

1.1 Overview

The modern information society is overloaded with a huge amount of data. Various search engines such as Google, Bing, and Yahoo make it easier for mass amounts of people to access information according to their need. Still, it is the peoples' job to go through all the documents returned by typical search engines. Not all the documents contain relevant information. At the same time most of the documents are filled with redundant information. People need only concise, relevant and core information that is content filtered from the ginormous information sources. Information content is dynamic in nature. For example, modern news agencies provide news on various events. They also provide frequent information updates on the same events. The users of those newspapers are not interested in going through all the information provided by news agencies. They are only interested in the updated information on the same event or topic. Well updated and summarized information can satisfy peoples' need. Due to the rise of information technologies, the demand for fully automated summarization systems is increasing rapidly. It is a wonderful way to abridge a large amount of redundant information into a condensed form by choosing the salient sentences and removing redundancy. Automatic MDS is to extract core information content from the source text, and present the most important content to the user in a concised form (Mani, 2001). The important contents are textual units or groups of textual units which should be taken into great consideration in generating a coherent and salient summary. The summaries are of tremendous value to many types of people, such as CEOs, students, lawyers and journalists. The abstract of scientific articles, headlines of a newspaper and reviews of a book are some examples of summarized content.

The task of summarization is to present concise and relevant information from single or multiple documents. Mani and Maybury (1999) define summarization as, "the process of distilling the most important information from the source (or sources) to produce an abridged version for a particular user (or users) and task (or tasks)". As humans are good at understanding natural language and knowing the required topic for the end user, they can perform this task. It would be much harder for them if the number of input documents increases. Also, the human generated summarization task becomes infeasible if documents are dynamic and rapid in changing in nature.

Several factors are considered during summarization process such as usage, compression rate[1], and functionality. Based on number of input source texts, summaries can be divided into two main types: single document summary and multi-document summary. The former is focused mainly on the core concept of text while the latter is focused on various criteria such as relevancy, redundancy, coherency, and responsiveness. Multi-document summarization (MDS) systems, which provide compressed and clustered information from a large number of documents, has become an important focus in the field of automatic text summarization. A set of related documents and optionally a query are considered as inputs of such systems. In multi-document summarization, sources contain the same information spread across various documents. It is a challenging job to extract all core information without destroying the structure and order of the source text. The tasks of MDS systems are to identify similar and dissimilar information, determine important contents, remove redundancy, and cover different aspects of the information which are available in the supplied source documents. A summarization task can be divided into two categories based on intended users: Generic summarization and Query-focused summarization. A generic summary is made for a wider section of users. Here, saliency is a major concern. Query-focused summarization system is designed for a concentrated section of users. Here, the

[1]Ratio of the summary length to the source text length.

sentence selection is expected to be biased towards the pieces of information closely related to the supplied queries. It can be generated on a fixed topic, or answers several user questions. The queries or questions can take any form. For example, it can be a factual question like "Who is X?" or a complex type like "Describe the effect of Cyclone Sidr". A generic summary contains the core information of the source documents, while a query-focused summary contains the information that can answer the need for information expressed in a topic (Wan et al., 2007). A generic MDS should contain the core information of the documents while keeping the minimum redundancy. As compared with generic MDS, the challenge for the query-focused MDS is that a query focused summary is not only expected to give important information contained in the document set but is also expected to guarantee that the information is related to the given topic.

Information sources such as news agencies provide news on the same topics but users are only interested in the updated news or information. An update MDS system can provide a solution for this problem. The goal of the update summarization is to get a salient summary of the updated documents assuming that the user has read the earlier documents about the same topic. An update summary is good for the users who need only the most recent information or news.

Ultimate Research Assistant[2], iResearch Reporter[3], Newsblaster[4], NewsFeed Researcher[5], and JistWeb[6] are some recent useful tools for automatic summary generation. Those systems automatically collect, cluster, categorize, and summarize news from several internet sites (CNN, Reuters, Fox News, etc.) and search engines (Google, Yahoo, etc.) on a daily basis. The event-related clusters are grouped from many news sources and each cluster provides one summary.

[2]http://ultimate-research-assistant.com/
[3]http://www.iresearch-reporter.com/
[4]http://newsblaster.cs.columbia.edu
[5]http://newsfeedresearcher.com
[6]http://www.jastatechnologies.com/productList.html

1.2 Our Approaches

The goal of this thesis is to propose a strategy for multi-document summarization using semantic atomic events. We extract all of the atomic semantic events from the source documents. We then rank sentences based on the importance of events and named entities. For more clarity, we propose methods to perform the following tasks: generic multi-document summarization, query-focused summarization, and guided update summarization. Events are the central components of sentences. Topic models help to identify all of the important events in source documents. Hence, it leads to better summary generation by using a good scoring scheme. For the query-focused summarization task, we use the vector space model where matrix entries are calculated based on the importance of semantic events and named entities. For guided update summarization, we use topic titles to guide summaries and the events are ordered according to their time of occurrences. Then the final ranking system is devised based on several important features such as new terms, topic title terms, temporal position of sentences.

We evaluate our system generated summaries in comparison with the supplied human generated summaries. We compare our systems with several most recent summarization systems and find that our generic and guided update summarization systems outperform the state-of-the-art summarization systems. Our query-focused summarization system performs better than most of the existing query-focused systems and is also statistically similar to the state-of-the-art systems. To measure the quality of our system generated summaries, we perform manual evaluation. Our systems generate coherent, responsive summaries which are measured by manual evaluation scores. The National Institute of Standards and Technology (NIST) supplied several datasets for summarization challenges. The data sets for Document Understanding Conferences (DUCs) are DUC 2004[7], DUC 2006[8], DUC

[7] http://duc.nist.gov/duc2004/tasks.html
[8] http://duc.nist.gov/duc2006/tasks.html

2007[9] etc. The data sets for Text Analysis Conferences (TACs) are TAC 2008[10], TAC 2009[11], TAC 2010[12], TAC 2011[13]. We use DUC 2004, DUC 2007, TAC 2010, and TAC 2011 data sets for all of our experiments.

1.3 Contribution

This thesis mainly contributes to the domain of generic, query-focused, and update summarization in the following ways.

We design a sentence ranking scheme which gives us a state-of-the-art generic summarization system. For query-focused summary, we design a system using both the vector space model and the topic model. For designing our update summarization system, we first design a system to calculate hidden sentence occurrence time by using event-temporal relations across the natural language texts. Then, a feature-rich scoring system is designed for better ranking of sentences for the summary.

1.4 Thesis Outline

The rest of the chapters of this thesis are arranged in the following ways:

- Chapter 2 reviews background and related summarization work and the widely-known available tools on which our summarization systems rely.

- Chapter 3 addresses our approach for generating summaries which are generic in nature. We also describe our approach for query-focused summarization for answering

[9] http://duc.nist.gov/duc2007/tasks.html
[10] http://www.nist.gov/tac/2008/
[11] http://www.nist.gov/tac/2009/
[12] http://www.nist.gov/tac/2010/
[13] http://www.nist.gov/tac/2011/

complex questions.

- Chapter 4 describes our novel approaches to generate update summarization guided by topics as well as a semantic approach of finding sentence time.
- Chapter 5 consists of the summarized concept of our results. We also identify some future directions for various summarization systems including generic, query-focused, and guided update summarization.

Chapter 2

Background

2.1 Introduction

Automatic text summarization presents the topical concepts found in one or more source documents to the users in a concise form so that the users no longer need to read all of the source documents. Extractive text summarization is achieved by selecting a small subset of sentences from the text collection, which are merged to form the summary. The summary should be coherent and contain the most relevant information from the source documents without unnecessary repetitions. Many researchers in the world working in the summarization field are using different approaches to get the best results. In this chapter, we review the related work in the summarization field as well as all the related tools and mathematical models. We also briefly mention the evaluation methods for summary. We mention the previous works related to generic and query-focused summary in section 3.2 in details. In section 4.2, we mention most of the current previous works related to update summarization.

2.2 Summarization techniques

Edmundson (1969) shows that a computer can be used for a summarization task. Since then, many researchers are involved in designing new summarization systems which can be implemented in a computer system.

Latent Semantic Analysis (LSA) (Landauer and Dumais, 1997) is a method for extracting and representing the semantic aspect of words, sentences, or documents by statistical computations. It has been extensively used in the field of information retrieval (Berry and

Fierro, 1996). It represents a set of source documents as a matrix where the rows of the matrix are the words from source documents and the columns are the sentences. The size of the matrix depends on the size of the input text. The dimensionality of a matrix rises with the increase of the number of sentences and distinct words in the source text. A matrix with high dimension means it contains a lot of noise, which can be reduced by decreasing the dimension in a scientific way. Singular Value Decomposition (SVD) can be that scientific model. SVD is used to obtain the orthogonal representation of sentences in the natural language texts. If there are m terms and n documents in the dataset, then we get a $m \times n$ term-document matrix A. The SVD of matrix A can be defined as:

$$A = U\sigma V \tag{2.1}$$

Where $U = [u_{ij}]$ is an $m \times n$ unitary matrix whose columns are called $left\ singular\ vectors$. We can interpret U as topic matrix where each column represents one topic.
$\sigma = diagonal(\sigma_1, \sigma_2, \ldots, \sigma_n)$ is an $n \times n$ diagonal matrix. $V = [v_{ij}]$ is an $n \times n$ unitary matrix whose columns are called *right singular vectors*. According to Steinberger (2007), "SVD maps between m-dimensional space specified by the weighted term-document vectors and the r-dimensional singular vector space, where r is the rank of Matrix A. From NLP point of view, SVD derives the 'latent semantic structure' of the document represented by matrix A, that means a breakdown of the original document into r linearly-independent base vectors which express the main topics of the document". Gong and Liu's (2001) method is one of the first steps to use an LSA-based lexical approach for multi-document summarization. They choose the sentences as summary sentences based on the relative importance of the topics. They use the transpose of the right singular matrix V to get summary sentences. Each row of matrix V^T represents one summary sentence. All the sentences extracted as summary sentences which represent an individual topic are equally important. As a result,

there is a possibility of including unimportant sentences (Steinberger, 2007). On the other hand, Steinberger (2007) chooses the sentences whose vectorial representation in the matrix $B = \sigma^2 V^T$ has the largest length, instead of the sentences containing the highest index value for each topic.

Lexical chains are defined as a semantic relationship among several nodes (single or multiple words) in a natural language text. The idea is motivated by the coherence and cohesion of text. The structure of a document in terms of macro-level relations between clauses or sentences is the coherence (Halliday, 1978). The cohesion property ensures that text unity is based on text elements relation (Halliday and Hasan, 2014). Semantic relations based on WordNet (Miller and Fellbaum, 1998) such as synonyms and hyponyms can be used as factors for measuring the lexical cohesion of a text. Barzilay *et al.* (1997) first uses lexical chains in text summarization. Although many researchers (Silber and McCoy, 2002; Barzilay et al., 1997; Stokes, 2004) compute lexical chains by directly using WordNet, Kolla (2004) uses topical relations. Mihalcea and Moldovan (2001) extends the WordNet method which is used to identify all of the topical relations. Kolla (2004) uses lexical chains to cluster segments of text and finally extracts one sentence from each cluster as a summary sentence.

NLP communities use several topic models to extract core summary information from the texts such as Latent Dirichlet Allocation (LDA) and Hierarchical Latent Dirichlet Allocation (HLDA). LDA (Blei et al., 2003), which is a Bayesian multinomial mixture model is very popular "in text analysis due to their simplicity, usefulness in reducing the dimensionality of the data, and ability to produce interpretable and semantically coherent topics" (Mimno and McCallum, 2012). It provides a generative probabilistic model that describes how the documents in a corpus are constructed. It analyzes collections of documents, each of which is represented as a mixture of topics, where each topic is a probability distribution over words or terms. The topic variable in the LDA model is selected repeatedly

within each document, which then allows documents to be comprised of many topics. This property is unique when compared to other generative models. Although Probabilistic Latent Semantic Indexing (PLSI) (Hofmann, 1999) also models documents into multiple topics, LDA is safe from data overfitting while adopting unseen documents because it uses a hidden random variable. The key idea behind modelling a document set with a dirichlet distribution over topics is that a document consists of many overlapping topics. For example, within a corpus of documents about the University of Lethbridge, there will be individual papers that discuss the Mathematics and Computer Science department. There will likely be certain words that are used more frequently when discussing the Mathematics and Computer Science department than other departments on campus such as, calculus, algorithm, database, and artificial intelligence. Other departments, such as the Neuroscience department may have papers discussing topics that use words like nervous system, cell, and brain. Arora and Ravindran (2008) use LDA and SVD to generate generic summaries of documents. They use LDA to split documents into different topics and SVD to get an orthogonal representation of sentences. They consider all the terms as equally contributing factors.

2.3 Events and Temporal Information Processing

Over time, the importance of a temporally-aware system has risen tremendously. For example, for a query focused summary of a newswire text, if we want to know who the Prime Minister (PM) of the United Kingdom was in the March of 1994, and we only had documents that tell us about the PM from 1990 to 1997 then a temporally aware system will help to infer who the PM was in the March of 1994 as well. Events and their temporal information have great significance in semantic learning about an article. Pustejovsky et al. (2003a) described the importance of events and their temporal relations in the following

ways:

"Events in articles are naturally anchored in time within the narrative of a text. For this reason, temporally grounded events are the very foundation from which we reason about how the world changes. Without a robust ability to identify and extract events and their temporal anchoring from a text, the real aboutness of the article can be missed."

Temporal annotation of a document is to extract all the events and their temporal relations as well as dates and times. TimeBank (Pustejovsky et al., 2003b) was the first complete temporally annotated corpus. It consists of around 200 documents, which are collected from various newswire texts. All the documents are annotated according to the TimeML standard (Pustejovsky et al., 2003a). For example, we can consider the following sentence taken from D1110B document set of TAC 2011 corpus:
"The Belgian government pledged on Tuesday an initial relief fund of 250,000 euros to China after a powerful earthquake struck southwestern China on Monday". The temporal annotation of the sentence would be:

"

<?xml version="1.0" encoding="UTF-8"?>

<TimeML >

<TIMEX3 tid="t0" functionInDocument="CREATION-TIME" type="DATE" value="2008-05-13"/ >

*The Belgian government <**EVENT eid="e1" class="I-ACTION" tense="PAST" aspect="NONE" polarity="POS" modality="none"** > pledged <**/EVENT** > on <**TIMEX3 tid="t1" type="DATE" value="2008-05-13"** > Tuesday <**/TIMEX3** > an initial relief fund of 250,000 euros to China after a powerful earthquake <**EVENT eid="e2" class="OCCURRENCE" tense="PAST" aspect="NONE" polarity="POS" modality="none"** > struck <**/EVENT**> southwestern China on <**TIMEX3 tid="t2" type="DATE" value="2008-05-12"** > Monday <**/TIMEX3**>.*

<TLINK relType="BEFORE" eventID="e1" relatedToTime="t0"/>

<TLINK relType="IS-INCLUDED" eventID="e1" relatedToTime="t1"/>

<TLINK relType="IS-INCLUDED" eventID="e2" relatedToTime="t2"/>

<TLINK relType="BEFORE" eventID="e2" relatedToTime="t0"/>

</TimeML>".

TIMEX3 and EVENT tags are used to represent times and events, respectively. TLINK tags are used to represent event-event and event-time relations. Various techniques are used to extract events and their temporal relations. The first complete event extraction called Evita, was developed by Saurí et al. (Saurí et al., 2005) whcih used "a linguistically motivated and rule based algorithm" to design their system. Bethard and Martin (2006) formulated the event extraction problem as a classification problem. They used a Support Vector Machine (SVM) with some linguistic features. In update summarization, knowing the relative order of the events is very useful to merge and present information from various news sources (Mani et al., 2003). In focused summarization (Pustejovsky et al., 2002), to infer the temporal ordering of events requires capabilities such as event occurrence time or which events occurred prior to a particular event (Mani et al., 2003). In addition, inferring relations of temporal entities and events is a crucial step towards the update summarization task. After the contribution of Pustejovsky et al. (2003a) , which is a TimeBank corpus, there have been several efforts to build temporally-aware systems and to find a hidden chronological order of events in corpora. Mani et al. (2003) was one of the first to infer hidden temporal relations of events. They showed that word and phrase level semantic information with the combination of sentence level syntactic information can anchor and order temporal events in a domain-independent corpus. Chambers and Jurafsky (2008) and Bramsen *et al.* (2006) consider only a few precedence relations such as *before* or *after* at the expense of expressiveness (Denis and Muller, 2011). There are a total of 12 relations in standard TimeML annotation (Pustejovsky et al., 2005). They are 'identical',

'before', 'ibefore', 'begin', 'end', 'include' and the rest are the inverses of the mentioned six relations. Allen (1983) and Denis and Muller (2011) take into account all possible temporal relations between temporal entities to make further reasoning possible. To reduce the combinatorial complexity of an ordering problem, Denis and Muller (2011) represent temporal interval constraints to end points and formulate the problem as an optimization problem. Yoshikawa et al. (2009) use Markov Logic Networks (Richardson and Domingos, 2006) for event-event association and their absolute time of occurrence by applying some inference rules. Correctly identifying associations between events and their absolute time of occurrence tasks are first addressed by Do et al. (2012). They use the Integer Linear Programming (ILP) formulation to combine a collection of local pairwise classifiers to get a universal timeline representation of events mapping with several lexical, syntactic and semantic features.

2.4 Summarization Challenges

The National Institute of Standards and Technology (NIST)[1] has been organizing an annual summarizing competition since 2001. Each year, they focus on specific summarization problems, distribute a dataset and mention evaluation rules and criteria. An expert is assigned to create and describe the problem. He/She also writes "topic", "topic title" and "narrative description" to help researchers focus on the summary. Document Understanding Conference (DUC) 2004 define five summarization tasks. Out of five tasks, task 2 is to generate short multi-document summaries for each data cluster where each is no more than 665 bytes (alphanumeric, white space, and punctuation included). Newswire documents are used by DUC 2004 challenge as the corpus. DUC 2007 has two types of tasks: main task and update task. The main task is to model questions which are complex in nature.

[1] http://www.nist.gov/

Stating only simple answers such as name, date or quantity will not answer those types of questions. There are 45 data clusters provided for that challenge. Each cluster contains 25 relevant documents and a set of complex questions are supplied for each cluster. The job is to generate a 250-word well-organized and fluent summary for each cluster that answers the complex question(s). The Text Analysis Conference (TAC) includes DUC as a Summarization track in 2008. A specific subproblem of the Natural Language Processing field is focused in each track of TAC. The TAC 2011 also consists of a summarization challenge track. Guided update summarization is a subproblem of that track. The task is to generate a 100-word update summary assuming that the user has already read the earlier documents of the same topic. We mainly focus on the tasks which we mention here.

Humans have a better analyzing and inferencing power than automatic systems do. So manual evaluation is a better way to measure quality, readability, and fluency of the system generated summaries. Various criteria are used for manual evaluation. For example, DUC 2007 manual evaluation consists of focus, non-redundancy, and overall responsiveness factors. A score is assigned from 1 (very poor) to 5 (very good) for each criterion.

Manual evaluation is costly and very slow. For rapid evaluation of a summary, manual evaluation is not realistic. A more immediate measurement process is needed to speed up comparison between systems. Recall-Oriented Understudy for Gisting Evaluation (ROUGE) (Lin, 2004) can fulfil that demand. ROUGE scores are measured by co-occurrence[2] between a system generated summary and a reference summary. There are many variation of ROUGE based on co-occurrence such as ROUGE-N, ROUGE-L, and ROUGE-SU. ROUGE-N is the n-gram co-occurrence measurement between two supplied summaries. *N* can be any number such as 1,2,3, and 4. N=1 means unigram co-occurrence, N=2 means bigram co-occurrence.

According to Jurafsky and Martin (2002), Recall (R) is a measurement metric which

[2]Co-occurrence means overlapping unit such as word sequence, n-gram, etc

measures the content in a candidate summary compared to a gold/reference summary. Precision (P) measures the quantity of a gold summary content in a candidate summary. F-measure is overall performance measurement system which combines both recall and precision. Unigram F-measure performance of a summary can be denoted by notation f_1.

2.5 Conclusion

In this chapter, we summarized various tools and techniques used in text summarization. Events and temporal relations processing were described briefly. The summary evaluation techniques were also explained for reader convenience. In the next chapter, we are going to explain our generic and query-focused summarization techniques. Our systems are compared with some other recent systems on the ROUGE metric. The overall quality of the generated summaries are measured manually.

Chapter 3

Generic and Query-focused Summarization

3.1 Introduction

In this chapter, we propose event based models for both the generic and query-focused MDS where we try to represent the summarization problem into an atomic event extraction as well as a topic distribution problem. The rest of the chapter consists of four sections. In section 3.2, we will look at some previous related works. Section 3.3 describes the generic MDS method. Section 3.4 gives the evaluation of our generic MDS system. Section 3.5 and 3.6 describe the query-focused MDS system and evaluation, respectively. Section 3.7 concludes this chapter.

3.2 Related Works

Every document covers a central theme or event. There are other sub-events which support the central event. There are also many words or terms across the whole document which can act as an individual event, they contribute to the main theme. Named entities such as time, date, person, money, organizations, locations, etc also have great significance of building up the structure of the document. Although events and named entities are terms or a group of terms, they have a higher significance than normal words or terms. Those events and named entities can help to generate high performing summaries. Filatova and Hatzivassiloglou (2004) use events in extractive summarization. They consider events as a triplet of two named entities and verb (or action noun) where verb (or action noun) connects the two named entities. Several greedy algorithms based on co-occurrence statistics of atomic events are used to generate a summary. They show that event-based summaries get

a much better score than the summaries generated by 'term frequency-inverse document frequency' (*tf***idf*) weighing of words. Li et al. (2006) also define the same complex structure as an event and the PageRank algorithm (Page et al., 1999) is applied to estimate the event relevance in summary generation. Another recent summarization work based on event semantics is done by Zhang *et al*. (2010). Their events may contain an unlimited number of entities. Due to the complex nature of all previous authors' defined events, it is hard to use their defined event concept in a topic model to get the semantic event distribution in text.

Our defined semantic event is an atomic term which is similar to the TimeML (Pustejovsky et al., 2003a). Pustejovsky et al. (2003a) consider events as a cover term for situations that happen, occur, hold, or take place. Event spans can be a period of time. Aspect, intentional state, intentional action, perception, occurrence, and modal can be events. The following are some examples of event expressions;

- Verbs: has arrested, will resigned, won;
- Event nominals: Vietnam War, Military operation;

Haghighi and Vanderwende (2009) show that HIERSUM which is a HLDA-style topic model (Blei et al., 2003) produces quality summaries. They consider all the words or terms as equally contributing factors in topic vocabulary distribution while we give the events and named entities high priority in content discovery. Event like deverbal nouns are used in G-FLOW (Christensen et al., 2013) to get discourse relations to ensure coherency in a summary. Our MDS systems use the event and entity distribution obtained from a topic model in sentence ranking to generate a quality summary.

3.3 Our Methodologies

3.3.1 Pre-processing of the data set

Pre-processing of multiple documents plays an important role for improving the recall and precision of a summary. We use the DUC 2004 and DUC 2007 datasets for generic and query-focused summarization, respectively. We use the Stanford CoreNLP[1] for sentence splitting, tokenization, named entity recognition, and cross-document coreference resolution. Stanford CoreNLP extracts all the coreference groups with their coreference mentions from the source text. We replace the coreference mentions with the coreference representative of that coreference group. For example, coreference resolution engine extracts "Cuban President Fidel Castro" as a representative of one coreference group. Then "Cuban President Fidel Castro" is used for mentions such as "Cuban President", "Fidel", "the President" and pronouns ("he", etc.) that the Stanford coreference resolution engine referred to "Cuban President Fidel Castro". New expanded documents sentences are used to feed the summarization systems which increase coverage and coherence of our summaries. We use the sentences from the original source dataset rather than the coreference resolved one for the final summary generation. We do it to get rid of error of the sentence structure and to preserve readability. To improve the summary recall, we remove all the candidate sentences containing quotations. We also remove candidate sentences whose length are less than 11 words. Sentences containing quotations are not appropriate for summary sentence and shorter sentences carry a small amount of information relatively (Li et al., 2013a). After tokenization, we remove stop words. We use the Porter Stemmer (Porter, 1980) for stemming. We use the ClearTK[2] system (Bethard, 2013) for event extraction.

[1] http://nlp.stanford.edu/software/corenlp.shtml
[2] http://code.google.com/p/cleartk/

3.3.2 Generic Summarization

Any cluster c contains d documents and all the documents are equiprobable, e.g. $P(D_{d'}) = 1$ where $P(D_{d'})$ is the probability of any document $D_{d'}$. All the documents in one cluster are sorted in descending order of their creation time[3]. The topic probability of each topic T_j can be calculated by equation (3.1) where $j \in \{1, \ldots, K\}$ and K is the number of topics for the Latent Dirichlet Allocation (LDA) Model. Probability of a term ($P(t)$) for any topic T_j can be computed by equation (3.2).

$$P(T_j) = \sum_{d'=1}^{d} P(T_j|D_{d'})P(D_{d'}) \tag{3.1}$$

$$P(t) = P(t|T_j)P(T_j) \tag{3.2}$$

To increase the coherency of the summary we calculate a sentence position score, S_p. If D_{ts} is the number of sentences in Document D, S_p can be calculated by equation (3.3) where sentence position index, $i \in \{0, \ldots, D_{ts} - 1\}$.

$$S_p = 1 - \frac{i}{D_{ts}} \tag{3.3}$$

The score of a sentence can be computed by equation (3.4).

$$Score(S) = S_p \times \sum_{t \in S} (P(t) \times W_g) \tag{3.4}$$

$$W_g = \frac{M}{TC_g} \tag{3.5}$$

[3]Document Creation Time (DCT) can be calculated from document name.

$$M = max_g TC_g, g \in \{e, n, o\} \tag{3.6}$$

In equation (3.4), W_g is the specific weight factor for each group of terms. TC_g is the number of terms in one group g where $g \in \{event(e), named\text{-}entity(n), other(o)\}$. We consider W_g is 1 for the group called *other* (which is set of normal terms other than events and named entities) and W_g for groups event and named entity can be calculated by equation (3.5).
If an event or named-entity group term count is the maximum, then the average weight of that group for all clusters is considered as its weight. Our weight calculating scheme ensures more weight for event and named entity groups and also prevents the high occurring group to get a higer score.

The steps of our generic summarization algorithm are mentioned below:

1. Apply the LDA Topic Model on the Corpus of Documents for a fixed number[4] of Topics K.

2. Compute the probability of Topic T_j by equation (3.1) and sort the topics in the descending order of their probabilities.

3. Pick the Topic T_j from sorted list where $P(T_1), \ldots, P(T_k)$ are in a sorted order.

4. For Topic T_j, compute the score of all sentences by equation (3.4) where $P(t)$ is the unigram probability distribution obtained from LDA topic model.

5. For Topic T_j, pick the sentence with highest score and include it in the summary. If it is already included in the summary or it dissatisfies other requirements (cosine score between candidate sentence and any one of the already-included summary sentence crosses the certain range), then pick the sentence with next highest score for this Topic T_j.

[4]Total 4 topics are taken into account, i.e. K is 4

6. Each selected sentence is compressed according to the method described in section 3.3.3.

7. If the summary reaches its desired length then terminate the operation, else continue from step 3.

3.3.3 Sentence Compression

Summary quality can be improved by sentence compression (Gillick et al., 2009; Li et al., 2011a). Suppose, a sentence *"The Amish school where a gunman shot 10 girls last week, killing five of them, is expected to be demolished on Thursday, a fire department official said."*. Here we can see the subclause "a fire department official said" does not have any significance in a summary. Removing those type of long unnecessary subclauses will improve summary quality as well as provide extra space to include new information in a fixed length summary. We mainly consider widely used reporting verbs such as said, told, or reported, to find out subclauses like the above example. We use the Stanford dependency parser (Cer et al., 2010) to parse the sentence. The dependency tree generated by the Stanford dependency parser is given below:

"

<dependencies type="collapsed-dependencies" >
<dep type="root">
<governor idx="0">ROOT</governor><dependent idx="29">said</dependent>
</dep><dep type="det">
<governor idx="3">school</governor><dependent idx="1">The</dependent>
</dep><dep type="amod">
<governor idx="3">school</governor><dependent idx="2">Amish</dependent>

</dep><dep type="tmod">

<governor idx="29">said</governor><dependent idx="3">school</dependent>

</dep><dep type="advmod">

<governor idx="19">expected</governor><dependent idx="4">where</dependent>

</dep><dep type="det">

<governor idx="7">shot</governor><dependent idx="5">a</dependent>

</dep><dep type="nn">

<governor idx="7">shot</governor><dependent idx="6">gunman</dependent>

</dep><dep type="nsubjpass">

<governor idx="19">expected</governor><dependent idx="7">shot</dependent>

</dep><dep type="num">

<governor idx="9">girls</governor><dependent idx="8">10</dependent>

</dep><dep type="dep">

<governor idx="7">shot</governor><dependent idx="9">girls</dependent>

</dep><dep type="amod">

<governor idx="11">week</governor><dependent idx="10">last</dependent>

</dep><dep type="tmod">

<governor idx="9">girls</governor><dependent idx="11">week</dependent>

</dep><dep type="partmod">

<governor idx="19">expected</governor><dependent idx="13">killing</dependent>

</dep><dep type="dobj">

<governor idx="13">killing</governor><dependent idx="14">five</dependent>

</dep><dep type="prep-of">

<governor idx="14">five</governor><dependent idx="16">them</dependent>

</dep><dep type="auxpass">

<governor idx="19">expected</governo><dependent idx="18">is</dependent>

</dep><dep type=“rcmod”>
<governor idx=“3”>school</governor><dependent idx=“19”>expected</dependent>
</dep><dep type=“aux”>
<governor idx=“22”>demolished</governor><dependent idx=“20”>to</dependent>
</dep><dep type=“auxpass”>
<governor idx=“22”>demolished</governor><dependent idx=“21”>be</dependent>
</dep><dep type=“xcomp”>
<governor idx=“19”>expected</governor><dependent idx=“22”>demolished</dependent>
</dep><dep type=“tmod”>
<governor idx=“22”>demolished</governor><dependent idx=“23”>Thursday</dependent>
</dep><dep type=“det”>
<governor idx=“28”>official</governo><dependent idx=“25”>a</dependent>
</dep><dep type=“nn”>
<governor idx=“28”>official</governor><dependent idx=“26”>fire</dependent>
</dep><dep type=“nn”>
<governor idx=“28”>official</governor><dependent idx=“27”>department</dependent>
</dep><dep type=“nsubj”>
<governor idx=“2”>said</governor><dependent idx=“28”>official</dependent>
</dep></dependencies>
”

Sentences containing a reporting verb are always being parsed following a fixed rule where the reporting verb is always the ‘root’ of the dependency tree. Then we traverse the parse tree to find out the subclause related to that reporting verb.

3.4 Evaluation: Generic Summarization

3.4.1 ROUGE Evaluation: Generic Summarization

To evaluate our generic MDS system we used the DUC 2004 dataset. We perform our experiment on 35 clusters of documents. Each cluster has 10 documents. The DUC 2004 Task-2 was to create short multi-document summaries no longer than 665 bytes. We evaluate our system generated summaries using the automatic evaluation toolkit ROUGE[5] (Lin, 2004). We compare our system with some recent good systems including the best system in DUC 2004 (Peer 65), conceptual units based model (which is augmented) (Takamura and Okumura, 2009), and G-FLOW, a recent state-of-the-art coherent summarization system (Christensen et al., 2013). As shown in Table 3.1, Our generic MDS system significantly outperforms those three systems. It also gets a better score than a recent submodular functions based state-of-the-art system[6] (Lin and Bilmes, 2011) based on ROUGE-1 recall and f_1-measure. For the DUC 2004 dataset, on average we find weight factors of the groups like events, named-entity, and others are 3, 1.14, and 1 respectively. That means our summarization system gives the highest priority to events group and less priority to normal terms. Hence, it explains the importance of semantic events in a summarization system. It also explains the importance of named entities over other tokens for summary generation.

Other recent summarization systems consider all terms/words as equally contributing factors for extracting important concept of the document. On contrary, we consider the events and named entities are the most important parts of the documents. Hence, we get better ROUGE scores compared to other state-of-the-art summarization systems.

[5]ROUGE runtime arguments for DUC 2004:
ROUGE -a -c 95 *-b* 665 *-m -n* 4 *-w* 1.2

[6]We do not compare our system with recent topic model based system (Haghighi and Vanderwende, 2009) because that system is significantly outperformed by Lin and Bilmes's (2011) system in both terms of ROUGE-1 recall and f_1-measure.

Systems	*R*-1	F_1
Peer 65	0.3828	0.3794
Takamura and Okumura (2009)	0.3850	-
G-FLOW	0.3733	0.3743
Lin	0.3935	0.3890
Our generic MDS system	**0.3953**	**0.3983**

Table 3.1: Evaluation on the DUC 2004 Dataset (best result is **bolded**)

Relevancy	3.92
Coherency	3.98
Non-redundancy	3.50
Overall responsiveness	3.70

Table 3.2: Manual evaluation on the DUC 2004 Dataset

3.4.2 Manual Evaluation: Generic Summarization

Human evaluation is necessary to get an accurate score of quality. We use the following criteria to manually evaluate our summary.

1. Relevancy
2. Non-redundancy
3. Coherency
4. Overall responsiveness

We randomly select 24 clusters from the DUC 2004 dataset and a total of 3 assessors are assigned for evaluation purposes. Each assessor examines all 24 clusters' summaries and gives a score of 1 (Very Poor) to 5 (Very Good). Finally the average scores are calculated. Table 4.3 tabulates average scores of manual evaluation on DUC 2004 dataset.

Our event-based summarization system chooses high relevance sentences as summary sentences. Our sentence position feature combined with other novel features help in select-

ing highly coherent sentences for the summary. But cosine similarity checking did a poor job in removing redundancy. It has been shown that a highly coherent summary "would have redundancy to some extent" (Takamura and Okumura, 2009). It may be the reason why our summarization system does not perform well in checking redundancy.

3.5 Query-focused summarization

For query-focused summarization, we use a Vector Space Model (VSM) (Salton et al., 1975) with LSA approximation. We use LDA to segregate the topics of the documents and VSM to relevance ranking with respect to the query terms. VSM is a document indexing model where index terms try to maintain a distant relation among each other in a document space. It is used in information retrieval, filtering, indexing and relevance rankings (Dubin, 2004). We can represent a collection of documents into a term-document matrix A where each entry (i,j) represents the importance of the term i in document j. The relevant information can be extracted based on query terms by taking the product $q^T A$ where **q** is the query vector. Similarly, we can calculate the score of the sentences instead of the documents from the term-sentence matrix. Same as generic summarization, we breakdown the terms of the LDA model into groups such as events, named entities, and others. Here we have to form the query vector from query title and narrative. We then process topic description file. The titles and narratives are extracted. They are tokenized and lemmatized. Then all the stop words are removed. We then use MIT Java Wordnet Interface[7] to get all linked words, mainly nouns, verbs, adjectives, and adverbs of remaining lemmatized terms. Then those terms are further stemmed[8]. These stemmed terms are used to form query vector by mapping terms to values formed by the LDA Model. We put some fixed values for titles

[7] http://projects.csail.mit.edu/jwi/

[8] Stemming is done after extracting linked words because Porter Stemmer does not always give us same token for all Parts-of-speech forms of a word.

and narratives in those index positions of query vector where stemmed query terms hit the LDA terms database. We use more weight for title terms than narrative terms.

For each topic T_j we can get the probability by equation (3.7) where $j \in \{1, \ldots, K\}$ and K is the number of topics for the LDA Model (Arora and Ravindran, 2008) and M is the number of documents in a dataset.

$$P(T_j) = \sum_{k=1}^{M} P(T_j|d_k)P(d_k) \tag{3.7}$$

In our case, each entry A_{ik} of the term-sentence matrix $A^{(j)}$ for the topic T_j can be calculated by the following formula.

$$A_{ik} = \begin{cases} W_g \times P(t_i|T_j)P(T_j|d_{d'})P(d_{d'}) & \text{if } t_i \in CS_k \\ 0 & \text{if } t_i \notin CS_k \end{cases}$$

In the above formula, CS_k represents the set of terms of the current sentence under consideration and W_g is the specific weight factor for each group of terms which can be calculated by equation (3.5). we use Vector Space Model (VSM) (Salton et al., 1975) with LSA-approximation. To the best of our knowledge, nobody uses combination of LDA and VSM (with LSA-approximation) Models for query-focused summarization task. Our contribution also lies on using atomic event rather than normal token. We follow same preprocessing step like generic summarization as described in subsection 3.3.1. Our query-focused summarization algorithm is as follows:

1. Pick the Topic T_j from sorted list where $P(T_1), \ldots, P(T_k)$ are in a sorted order.

2. Calculate term-sentence matrix A and construct its "Singular Value Decomposition" (SVD) where $A = U\sigma V^T$. (Please see section 2.2 for details)

3. We get U_k by reducing U's column to k, σ_k by reducing both its row and column to

k, and finally V_k by reducing its column to k where $k = \lfloor \sqrt{A_c} \rfloor$ and A_c is the column dimension of matrix A

4. Compute and output $A_k = U_k \sigma_k {V_k}^T$ as the rank-k approximation to A.

5. Compute the sentence score from $q^T A_k$ and sort the sentences in the descending order of their scores.

6. Similar to generic summarization choose sentences from every topic T_j as final summary sentences.

7. Like generic summarization, each selected sentence is compressed according to the method described in section 3.3.3.

3.6 Evaluation: Query-focused Summarization

3.6.1 ROUGE Evaluation: Query-focused Summarization

We have evaluated our query-focus summarization system for the DUC 2007 Main task. We perform our experiment on 28 clusters of document. The length of summary should be no longer than 250 words. Table 3.3 tabulates the ROUGE[9] scores of our system and the best performing systems in the DUC 2007 task. We also get an improved score for high weighted events, and named entities. DUC 2007 automatically evaluate all the submitted systems by giving priority to ROUGE-2 and ROUGE-SU4 scores. Our query-focused MDS system outperforms Two-tiered topic model (TTM) (Celikyilmaz and Hakkani-Tür, 2011), a state-of-the-art unsupervised query-focused summarization system in terms of ROUGE-2 recall measure. Our system does not outperform the best DUC 2007 system which is

[9] *ROUGE* runtime arguments for DUC 2007:
ROUGE -n 2 *-x-m* -2 4 *-u -c* 95 *-r* 1000 *-f A-p* 0.5*-t* 0*-d*

System	ROUGE-2	ROUGE-SU4
Baseline	0.06039	0.10507
Our Query-focused MDS system	0.11183	0.16427
Tree (HEAD + Multi)	**0.1349**	**0.1846**
TTM	10.7	-
MNSF	12.38	-
Best participating DUC 2007 system	0.12448	0.17711

Table 3.3: ROUGE evaluation on the DUC 2007 Dataset (best result is **bolded**)

not fully unsupervised and uses the Yahoo search engine to get a ranked set of retrieved documents from the web with the DUC 2007 topic title being the query (Pingali and Varma, 2007). All of the state-of-the-art query-focused summarization systems such as parse-tree-based sentence compression system, which use head-driven beam search decoder system (e.g. Tree (HEAD + Multi)) (Wang et al., 2013). Monotone nondecreasing submodular functions (MNSF) based system (Lin and Bilmes, 2011) are fully supervised. However, our system is completely unsupervised. Our query-focused system's ROUGE-2 score is statistically significant compared to **TTM** (Celikyilmaz and Hakkani-Tür, 2011) based on t-test on 95% confidence level.

3.6.2 Manual Evaluation: Query-focused Summarization

Like generic summarization, we perform a manual evaluation for our query-focused summarization system. Like the DUC 2007 summarization challenge, we use focus, non-redundancy, and overall responsiveness as evaluation criteria. We randomly select 21 clusters from the DUC 2007 dataset and a total of 3 assessors are assigned for evaluation pur-

Focus	3.70
Non-redundancy	3.63
Overall responsiveness	3.85

Table 3.4: Manual evaluation on the DUC 2007 Dataset

poses. Each assessor examines all 21 cluster summaries and gives a score of 1 (Very Poor) to 5 (Very Good). Finally, the average scores are calculated. Table 3.4 tabulates average scores of manual evaluation on DUC 2007 dataset.

3.7 Conclusion

In this chapter, we showed a simple yet effective way of approaching the task of generating generic summaries. The importance of semantic events and named entities in generating summaries has been deeply analyzed using the LDA topic model. By dividing terms into different groups we achieve high ROUGE scores for the generic MDS task. Our query-focused summarization system shows a statistically similar result to the state-of-the-art unsupervised query-focused summarization system.

The Guided update summarization task, which is a new type of challenge for summarization communities is proposed by NIST[10] to "encourage a deeper linguistic (semantic) analysis of the source documents instead of relying only on document word frequencies to select important concepts. It is guided in the sense that the generated summary topic falls into a predefined category. Participants are given a list of aspects for each category, and a summary must include all aspects found for its category". The job is to generate a short 100-word guided update summary of newswire texts related to a topic, "under the assumption that the user has already read the earlier articles of the same topic"[11]. In the next chapter, we describe our guided update summarization system which is designed based on semantic events and temporal

[10] www.nist.gov/tac/

[11] http://www.nist.gov/tac/2011/Summarization/

relations. The system results are also compared with other recent systems using modern evaluation criteria.

Chapter 4

Guided Update Summarization

4.1 Introduction

In this chapter, we design a novel approach that takes into account all the events in a sentence and their temporal relations to ensure its novelty as well as its saliency in update summarization. We represent the novelty detection problem as a chronological ordering problem of the temporal events and time expressions. Our event based sentence ranking system uses a topic model that identifies all of the salient sentences. The rest of the chapter consists of four sections. In section 4.2, we will look at some previous works related to update MDS. Section 4.3 describes our guided update MDS system. Section 4.4 gives the evaluation of our system. Section 4.5 concludes this chapter.

4.2 Related Works

Update summarization, the newest type of challenge for summarization communities is introduced first in DUC 2007[1]. The update summarization task is very challenging because it is a problem of novelty detection without losing saliency and coherency. The most recent efforts to generate an update summary use graph based algorithms with some additional features to explore the novelty of the documents (Wenjie et al., 2008; Du et al., 2010; Li et al., 2011b). The Maximal Marginal Relevance (MMR) based approach (Boudin and El-Bèze, 2008) is used to blindly filter out the new information. These approaches discard the sentences containing novel information if they contain some old information from the previous document sets (Delort and Alfonseca, 2012).

[1] http://duc.nist.gov/duc2007/tasks.html

Steinberger *et al.* (2011) use the sentence time information in the Latent Semantic Analysis (LSA) framework to get the novel sentences. They only consider the first time expression as the anchored time of the sentence, but sentences may contain multiple time expressions from various chronologies. For instance, consider the sentence *"Two members of Basque separatist group ETA arrested while transporting half a tonne of explosives to Madrid just prior to the* ***March 2004*** *bombings received jail sentences of 22 years each on* ***Monday****."*. Here we get two[2] time expressions, which are *March 2004* and *Monday*. The first expression represents the very old information and the second one represents the new information. If we consider the first time expression as the sentence time like Steinberger *et al*. (2011), then it would give us false novel/update information. This is why we take into account all the events of a sentence to calculate its anchored time.

Lin et al. (2008) first use the temporal relations among events in update summarization. They try to find out the event-event links of the adjacent sentences. The precedence of a sentence is decided based on only one event of that sentence. In many cases, a sentence contains multiple events. Several popular generic summarization approaches such as LexRank (Erkan and Radev, 2004), TextRank (Mihalcea and Tarau, 2004) were used in update summarization without paying attention to the novelty detection. Fisher and Roark (2008) used a domain-independent supervised classification to rank sentences and extract all the sentences containing old information by using some filtering rules. By formulating the problem as an Integer Linear Programming (ILP) problem, high ROUGE scores are obtained for TAC 2009 (Conroy et al., 2009; Gillick et al., 2009). While the former one tries to maximize the sum of oracle scores, the latter one extracts sentences which maximize the sum of weights of n-grams. Some recent approaches (Cheng et al., 2013; Li et al., 2013b) use a reinforcement process to ensure the saliency in update summarization. Cheng et al. (2013) consider the old documents as sink points and penalize the sentences

[2]Here **'22 years'** is a time period. Time periods do not carry important information for detecting novelty.

sharing similar information with the old documents based on manifold ranking (Zhou et al., 2003). In 'Positive and Negative Reinforcement'(PNR) by Li et al. (2008) and 'Manifold Ranking with Sink Points' (MRSP) by Du et al. (2010), old documents act as constraints in reinforcement propagation. QCQPSum[3] (Li et al., 2013b) involves the previous documents in both the objective function formulation and the reinforcement propagation in new documents. They do not try to extract the novel information at a semantic level. We can see a few semantic analysis-based novelty detection approaches. Those are a Iterative Residual Rescaling (IRR) based LSA framework (Steinberger and Ježek, 2009) and a Bayesian multinomial probability distribution based approach (Delort and Alfonseca, 2012). The state-of-the-art update summarization system, the h-uHDP model (Li et al., 2012) uses a Hierarchical Dirichlet Process (HDP) (Teh et al., 2004) to get the history epoch and the update epoch distribution. They use the Kullback-Leibler (KL) (Kullback, 1987) divergence based greedy approach to select novel sentences. They all neglect semantic temporal information which is crucial in novelty detection. We use the TAC 2011 dataset for update summarization evaluation. Similar to generic and query-focused summarization, we use the same pre-processing technique.

4.3 Our Methodologies

4.3.1 Time End Point Normalization

Time expression identification and normalization are integral parts for the temporal processing of the raw text. We use the Stanford SUTime (Chang and Manning, 2012) which is a rule-based temporal tagger extracts all the temporal expressions. SUTime is one of the best systems for capturing temporal expressions from a natural language text. It follows

[3]"Quadratically constrained quadratic programming" (QCQP) is an optimization problem and NP-hard.

TimeML (Pustejovsky et al., 2003a) formats (TIMEX3) for normalizing time expressions. Consider the sentence: *"The Amish school where a gunman shot 10 girls* ***last week****, killing five of them, is expected to be demolished* ***Thursday****, a fire department official said."* Here *last week* and *Thursday* are the time expressions of the sentence. The SUTime output of the above text is mentioned below where October 11^{th}, 2006 is the reference date:

"The Amish school where a gunman shot 10 girls <TIMEX3 tid="t2" type="DATE" ***value="2006-W40"****>last week</TIMEX3>, killing five of them, is expected to be demolished <TIMEX3 tid="t3" type="DATE"* ***value="2006-10-12"*** *>Thursday</TIMEX3 >, a fire department official said."* SUTime extracts *2006-W40* and *2006-10-12* as the normalized date of *last week* and *Thursday* respectively. We convert them into an absolute time end point on a universal timeline. We follow standard date and time format (*YYYY-MM-DD hh:mm:ss*) for the time end point. For example, After conversion of **2006-W40** and **2006-10-12**, we get 2006-09-23 23:59:59 and 2006-10-12 23:59:59 respectively.

4.3.2 Temporal Ordering of Events and Time Expressions

An Event is an action or occurrence that happens with associated participants or arguments (Do et al., 2012). Filatova and Hatzivassiloglou (2004) describe an event as a major constituent part of action described in a narrative. They use an event as a feature for an extractive summarization task. They define event as a triplet of two named entities connected by verb or an action-denoting noun. Our definition of an atomic event is smaller in size and covers all possible events that can be represented in a real world text. Capturing the important semantic atomic events means capturing the core concepts of the documents, which is the goal of the summarization task. We use the same event concept similar to our generic summarization system. Unlike Denis and Muller (2011), we anchor events to only one time point, which is the upper end point. We are only concerned about the relative

ordering of the events. TempEval challenges such as Verhagen et al. (2007), Verhagen et al. (2010), UzZaman et al. (2013) defined several temporal tasks. Those are:

1. Extracting all possible events and timex node in texts.
2. Determining the value and type of timex node defined by TIMEX3 tag.
3. Identifying temporal relation between main events of consecutive sentences.
4. Identifying temporal relation between pair of events in the same sentence.
5. Identifying the relations among events and timex nodes in the same sentence.
6. Identifying event and document creation time (DCT).

The tasks 3 to 6 are called temporal relation tasks. Out of the many participating systems in the 2013 TempEval task, ClearTK-TimeML (Bethard, 2013) ranked first for task-related to temporal relations. Therefore, we use the ClearTK-TimeML tool to extract events and temporal relations. In ClearTK-TimeML (Bethard, 2013), four types of temporal relations are predicted. They are **B**EFORE, **A**FTER, **I**NCLUDES, and **N**ORELATION. Our main goal is to solve the novelty problem by using relative events' anchored values. In order to saturate the event-event and event-time relations, we use several transitive closure rules (Allen, 1983; Setzer et al., 2003). Some of them are given below:

A before B and B before C $\implies$ *A before C*

A includes B and B includes C $\implies$ *A includes C*

A after B and B after C $\implies$ *A after C*

We anchored all the events to absolute times based on the 'includes' and 'is-included' relation of the event-time link. The remaining events are anchored approximately based on other relations which are *'before'*, and*'after'*.

4.3.3 Temporal Score

We use the ClearTK system (Bethard, 2013) for initial temporal relation extraction and use some transitive rules as described in subsection 4.3.2. We relax the original event-time association problem by anchoring an event to an approximate time. We calculate the temporal score of a sentence by taking an average time score of all event-anchored times. Then, all the sentences are ordered in descending order of their temporal scores except the first sentence of each document. Then we calculate the Temporal Position Score (tp_s) of the temporally ordered sentences. The tp_s of the first sentence of a document is considered one. The temporal position scores tp_s of other sentences (except the first sentence of each document) can be calculated by equation (4.1). D_s is the number of sentences in Document D and the temporally ordered sentence position index, $i \in \{0, \ldots, D_s - 1\}$.

$$tp_s = 1 - \frac{\gamma \times i}{D_s} \tag{4.1}$$

The parameter γ is used to tune the weight of relative temporal position of the sentences. The temporal distance between two consecutive sentences increases with the increase of the value of γ and decreases with the decrease of the value of γ.

4.3.4 Sentence Ranking

From the Latent Dirichlet Allocation (LDA) topic model[4] we get a unigram (event or named entity) probability distribution, P (t). For each topic, the sentence score can be computed by equation (4.2).

$$Score(S) = tp_s \times \sum_{t \in S} (P(t) \times (\alpha + \beta) \times W_g) \tag{4.2}$$

[4]Total 4 topics are taken into account, i.e. K is 4

Systems	*ROUGE*-2	*ROUGE-SU*4
Our System	**0.1074**	**0.1430**
h-uHDPSum	0.1017	0.1364
Peer 43	0.0959	0.1309

Table 4.1: Evaluation on the TAC 2011 Dataset (best result is **bolded**)

Systems	*ROUGE*-2	*ROUGE-SU*4
Our System	**0.1031-0.1204**	**0.1370-0.1551**
h-uHDPSum	0.0910-0.1034	0.1265-0.1473
Peer 43	0.0894-0.1029	0.1251-0.1366

Table 4.2: 95% confidence for various systems on the TAC 2011 Dataset (best result is **bolded**)

In equation (4.2), tp_s is the temporal position score of the sentence obtained from equation (4.1) and α and β are the weight factors of the new terms and the topic title terms, respectively, which are learnt empirically from TAC 2010 dataset. For each topic, one sentence is taken as a summary sentence from the ordered list of sentences based on descending order of their score ($Score(S)$). A Cosine similarity score is used to remove redundancy from the summary. Like the generic and query-focused summarization tasks, we use the same sentence compression technique.

4.4 Evaluation

4.4.1 ROUGE Evaluation

To evaluate our update MDS system we used the TAC 2011 dataset. TAC 2011 dataset contains two groups (A and B) of data. Group A contains the old dataset. Group B contains a new dataset of the same topic of group A. We perform our experiment on 28 clusters of documents. Each cluster has 10 documents. The TAC 2011 Guided update summarization task was to create short multi-document summaries no longer than 100 words with the

assumption that the user has already read documents from the group A dataset. We evaluate our system generated summaries using the automatic evaluation toolkit ROUGE[5] (Lin, 2004). Table 4.1 tabulates our system and best performing systems in the TAC 2011 update summarization task. Our model outperforms the current state-of-the-art system, which is h-uHDPSum, as well as the best update summarization system (peer 43) of TAC 2011 summarization track. The 95% confidence intervals in Table 4.2 show that our system obtains a statistically significant improvement over the peer 43 system in terms of ROUGE-2 and ROUGE-SU4. The performance of our event and temporal relation-based summarizer changes according to the type of documents we are considering to be summarized. Our system gets very high recall and f-measures for documents which are well defined constituents of events. Our temporal relation based system reveals all hidden novel information. At the same time our event and named entity-based scoring scheme ensures saliency in update summarization.

4.4.2 Manual Evaluation

A human evaluation is necessary to measure the quality of summary properly. We use the following criteria to manually evaluate our summary.

1. Novelty (containing update information)
2. Readability/Fluency
3. Overall responsiveness (Overall focus and content)

We randomly select 21 clusters from the TAC 2011 dataset, and a total of 3 assessors are assigned for evaluation purposes. Each assessor examines all 21 cluster summaries and

[5]ROUGE runtime arguments for TAC 2011:
ROUGE -l 100-*n* 4-*w* 1.2-*m* -2 4 *-u -r* 1000 *-f A -p* 0.5 *-t* 0-*a -d*

Novelty	4.13
Fluency	3.92
Overall responsiveness	4.07

Table 4.3: Manual evaluation on the TAC 2011 Dataset

gives a score of 1 (Very Poor) to 5 (Very Good). Finally, average scores are calculated. Table 4.3 tabulates the scores of manual evaluation on TAC 2011 dataset. Our temporal summarization system chooses high novel sentences as summary sentences without losing fluency and responsiveness.

4.5 Conclusion

In this chapter, we have shown the way to generate update summaries using semantic events and their temporal relations. "Temporally grounded events" can effectively extract updated information from the source text. Our sentence ranking system combines the power of topic distribution and semantic temporal relations to generate salient update summaries. In the next chapter, we will mention some future direction in summarization tasks and conclude the thesis.

Chapter 5

Conclusion and Future Work

In this thesis, we focus on solving various summarization problems in the context of semantic atomic events and their temporal relations. Our defined events are atomic which made it possible to integrate them into an LDA topic model to get the overall topic distribution of a source text. We have also shown that events and named entities are more important than normal words/terms in getting good summaries. Our generic MDS system outperforms all the recent state-of-the-art generic MDS systems in terms of ROUGE-1 recall measure. Our feature-rich generic sentence ranking technique ensures coherency in a summary.

We designed our query-focused MDS system by using events in LDA and VSM models. LDA topic model segregates documents into topics and VSM ensures relevancy in a summary. To the best of our knowledge, we are the first to use LDA topic model with VSM relevance ranking where matrix entries are calculated based on the importance of semantic events and named-entities. Our query-focused MDS system achieves a statistically similar result to the state-of-the-art query-focused MDS systems in terms of ROUGE-2 recall. Other topic models like the Pachinko Allocation Model (PAM) (Li and McCallum, 2006) and the Hierarchical Pachinko Allocation Model (HPAM) (Mimno et al., 2007) can be used instead of LDA to model correlation between topics along with word correlations. Their ability to support finer-grained topics and topical keyword coherence (Li and McCallum, 2006) can further improve sentence ranking. Dirichlet Multinomial Regression (DMR) is another promising topic model which includes a log-linear calculation prior to document-topic distributions (Mimno and McCallum, 2012). By selecting appropriate features, DMR can increase the recall of the summarization. It is worthy of effort to identify another contributing group of terms like event and named entity to acquire high recall and precision for query-focused summary.

Our update summarization model can identify novel information based on temporal ordering of events and times. Our scoring scheme using a topic model detects all the available topics in a corpora. Our system outperforms the state-of-the-art update summarization systems based on the ROUGE-2 and ROUGE-SU4 recall measures. However, event-event and event-time ordering can still be improved. Denis and Muller (2011) stated that ordering temporal entity considering for all possible 12 relations is NP-complete problem. They reduce the complexity of the problem by converting relations into end points. However, they get a 41% F1-score. By increasing the recall and precision of event-event and event-time relation extraction, it is possible to get better temporal ordering of sentences. That will eventually provide better update summarization. Some recent works on temporal relation classification of using dependency parses (Ng and Kan, 2012) and discourse analysis framework (Ng et al., 2013) can further improve our summarization system performance.

References

James F Allen. 1983. Maintaining knowledge about temporal intervals. *Communications of the ACM*, 26(11):832–843.

Rachit Arora and Balaraman Ravindran. 2008. Latent Dirichlet Allocation and Singular Value Decomposition based multi-document summarization. In *Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on*, pages 713–718. IEEE.

Regina Barzilay, Michael Elhadad, et al. 1997. Using Lexical Chains for Text Summarization. In *Proceedings of the Association for Computational Linguistics workshop on intelligent scalable text summarization*, volume 17, pages 10–17. Madrid, Spain.

Michael W Berry and Ricardo D Fierro. 1996. Low-rank orthogonal decompositions for information retrieval applications. *Numerical linear algebra with applications*, 3(4):301–327.

Steven Bethard and James H Martin. 2006. Identification of event mentions and their semantic class. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 146–154. Association for Computational Linguistics.

Steven Bethard. 2013. ClearTK-TimeML: A Minimalist Approach to TempEval 2013. In *Second Joint Conference on Lexical and Computational Semantics (* SEM)*, volume 2, pages 10–14.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent Dirichlet Allocation. *The Journal of Machine Learning Research*, 3:993–1022.

Florian Boudin and Marc El-Bèze. 2008. A Scalable MMR Approach to Sentence Scoring for Multi-document Update Summarization.

Philip Bramsen, Pawan Deshpande, Yoong Keok Lee, and Regina Barzilay. 2006. Inducing temporal graphs. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 189–198. Association for Computational Linguistics.

Asli Celikyilmaz and Dilek Hakkani-Tür. 2011. Discovery of Topically Coherent Sentences for Extractive Summarization. In *Association for Computational Linguistics*, pages 491–499.

Daniel M Cer, Marie-Catherine De Marneffe, Daniel Jurafsky, and Christopher D Manning. 2010. Parsing to Stanford Dependencies: Trade-offs between Speed and Accuracy. In *Language Resources and Evaluation (LREC 2010)*.

Nathanael Chambers and Dan Jurafsky. 2008. Jointly combining implicit constraints improves temporal ordering. In *Proceedings of the Conference on Empirical Methods in*

Natural Language Processing, pages 698–706. Association for Computational Linguistics.

Angel X Chang and Christopher D Manning. 2012. SUTIME: a library for recognizing and normalizing time expressions. *Language Resources and Evaluation (LREC 2012)*, pages 3735–3740.

Xue-Qi Cheng, Pan Du, Jiafeng Guo, Xiaofei Zhu, and Yixin Chen. 2013. Ranking on Data Manifold with Sink Points. *Knowledge and Data Engineering, IEEE Transactions on*, 25(1):177–191.

Janara Christensen, Mausam, Stephen Soderland, and Oren Etzioni. 2013. Towards Coherent Multi-document Summarization. In *Proceedings of NAACL-HLT*, pages 1163–1173.

John M Conroy, Judith D Schlesinger, and Dianne P Oleary. 2009. CLASSY 2009: Summarization and Metrics. In *Proceedings of the text analysis conference (TAC)*.

Jean-Yves Delort and Enrique Alfonseca. 2012. DUALSUM: a Topic-Model Based Approach for Update Summarization. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 214–223. Association for Computational Linguistics.

Pascal Denis and Philippe Muller. 2011. Predicting globally-coherent temporal structures from texts via endpoint inference and graph decomposition. In *Proceedings of the Twenty-Second international joint conference on Artificial Intelligence-Volume Three*, pages 1788–1793. AAAI Press.

Quang Xuan Do, Wei Lu, and Dan Roth. 2012. Joint inference for event timeline construction. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 677–687. Association for Computational Linguistics.

Pan Du, Jiafeng Guo, Jin Zhang, and Xueqi Cheng. 2010. Manifold Ranking with Sink Points for Update Summarization. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 1757–1760. Association for Computing Machinery.

David Dubin. 2004. The most influential Paper Gerard Salton never wrote. *Library trends*, 52(4):748–764.

Harold P Edmundson. 1969. New methods in automatic extracting. *Journal of the ACM (JACM)*, 16(2):264–285.

Günes Erkan and Dragomir R Radev. 2004. LexRank: Graph-based Lexical Centrality as Salience in Text Summarization. *J. Artif. Intell. Res.(JAIR)*, 22(1):457–479.

Elena Filatova and Vasileios Hatzivassiloglou. 2004. Event-based Extractive Summarization. In *Proceedings of Association for Computational Linguistics Workshop on Summarization*, volume 111.

Seeger Fisher and Brian Roark. 2008. Query-focused Supervised Sentence Ranking for Update Summaries. *Proceeding of Text Analysis Conference (TAC 2008)*.

Dan Gillick, Benoit Favre, Dilek Hakkani-Tur, Berndt Bohnet, Yang Liu, and Shasha Xie. 2009. The ICSI/UTD summarization system at TAC 2009. In *Proceedings of the Second Text Analysis Conference, Gaithersburg, Maryland, USA. National Institute of Standards and Technology*.

Yihong Gong and Xin Liu. 2001. Generic text summarization using relevance measure and latent semantic analysis. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 19–25. Association for Computing Machinery.

Aria Haghighi and Lucy Vanderwende. 2009. Exploring Content Models for Multi-document Summarization. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 362–370. Association for Computational Linguistics.

Michael Alexander Kirkwood Halliday and Ruqaiya Hasan. 2014. *Cohesion in English*. Routledge.

Michael AK Halliday. 1978. *Language as Social Semiotic: The Social Interpretation of Language and Meaning.* University Park Press.

Thomas Hofmann. 1999. Probabilistic Latent Semantic Indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 50–57. Association for Computing Machinery.

D. Jurafsky and J. H. Martin. 2002. *Speech and Language Processing*. Amazon, second edition.

Maheedhar Kolla. 2004. Automatic text summarization using lexical chains: Algorithms and experiments. Master's thesis, Lethbridge, Alta.: University of Lethbridge, Faculty of Arts and Science, 2004.

Solomon Kullback. 1987. The Kullback-Leibler distance. *American Statistician*, 41(4):340–340.

Thomas K Landauer and Susan T Dumais. 1997. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, 104(2):211.

Wei Li and Andrew McCallum. 2006. Pachinko allocation: DAG-structured mixture models of topic correlations. In *Proceedings of the 23rd international conference on Machine learning*, pages 577–584. Association for Computing Machinery.

Wenjie Li, Mingli Wu, Qin Lu, Wei Xu, and Chunfa Yuan. 2006. Extractive Summarization using Inter-and Intra-event Relevance. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 369–376. Association for Computational Linguistics.

Peng Li, Yinglin Wang, Wei Gao, and Jing Jiang. 2011a. Generating Aspect-oriented Multi-document Summarization with Event-aspect Model. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1137–1146. Association for Computational Linguistics.

Xuan Li, Liang Du, and Yi-Dong Shen. 2011b. Graph-Based Marginal Ranking for Update Summarization. In *SDM*, pages 486–497. Society for Industrial and Applied Mathematics (SIAM).

Jiwei Li, Sujian Li, Xun Wang, Ye Tian, and Baobao Chang. 2012. Update Summarization Using a Multi-level Hierarchical Dirichlet Process Model. In *COLING*, pages 1603–1618.

Lei Li, Wei Heng, Jia Yu, Yu Liu, and Shuhong Wan. 2013a. CIST System Report for ACL Multiling 2013–track 1: Multilingual Multi-document Summarization. *MultiLing 2013*, page 39.

Xuan Li, Liang Du, and Yi-Dong Shen. 2013b. Update Summarization via Graph-based Sentence Ranking. *Knowledge and Data Engineering, IEEE Transactions on*, 25(5):1162–1174.

Hui Lin and Jeff Bilmes. 2011. A Class of Submodular Functions for Document Summarization. In *Association for Computational Linguistics*, pages 510–520.

Ziheng Lin, Huu Hung Hoang, Long Qiu, Shiren Ye, and Min-Yen Kan. 2008. NUS at TAC 2008: Augmenting Timestamped Graphs with Event Information and Selectively Expanding Opinion Contexts. In *Proceedings of Text Analysis Conference (TAC 2008) Workshop on Automatic Summarization, Gaithersburg*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, pages 74–81.

Inderjeet Mani and Mark T Maybury. 1999. *Advances in Automatic Text Summarization*, volume 293. MIT Press.

Inderjeet Mani, Barry Schiffman, and Jianping Zhang. 2003. Inferring temporal ordering of events in news. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: companion volume of the Proceedings of HLT-NAACL 2003–short papers-Volume 2*, pages 55–57. Association for Computational Linguistics.

Inderjeet Mani. 2001. *Automatic summarization*, volume 3. John Benjamins Publishing.

Rada Mihalcea and Dan I Moldovan. 2001. Extended Wordnet: Progress Report. In *in Proceedings of NAACL Workshop on WordNet and Other Lexical Resources*. Citeseer.

Rada Mihalcea and Paul Tarau. 2004. TextRank: Bringing Order into Texts. In *Proceedings of EMNLP*, volume 4, page 275. Barcelona, Spain.

George Miller and Christiane Fellbaum. 1998. Wordnet: An Electronic Lexical Database.

David Mimno and Andrew McCallum. 2012. Topic models conditioned on arbitrary features with dirichlet-multinomial regression. *arXiv preprint arXiv:1206.3278*.

David Mimno, Wei Li, and Andrew McCallum. 2007. Mixtures of hierarchical topics with pachinko allocation. In *Proceedings of the 24th international conference on Machine learning*, pages 633–640. Association for Computing Machinery.

Jun-Ping Ng and Min-Yen Kan. 2012. Improved Temporal Relation Classification using Dependency Parses and Selective Crowdsourced Annotations. In *Proceedings of the 24th International Conference on Computational Linguistics*, pages 2109–2124.

Jun-Ping Ng, Min-Yen Kan, Ziheng Lin, Wei Feng, Bin Chen, Jian Su, and Chew Lim Tan. 2013. Exploiting Discourse Analysis for Article-Wide Temporal Classification. In *EMNLP*, pages 12–23. Association for Computational Linguistics.

Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. 1999. The PageRank Citation Ranking: Bringing Order to The Web.

RK Pingali and Vasudeva Varma. 2007. IIIT Hyderabad at DUC 2007. *Proceedings of DUC 2007*.

Martin F Porter. 1980. An algorithm for suffix stripping. *Program: electronic library and information systems*, 14(3):130–137.

James Pustejovsky, Janyce Wiebe, and Mark Maybury. 2002. Multiple-perspective and Temporal Question Answering. In *Question Answering: Strategy and Resources Workshop Program*, page 43.

James Pustejovsky, José M Castano, Robert Ingria, Roser Sauri, Robert J Gaizauskas, Andrea Setzer, Graham Katz, and Dragomir R Radev. 2003a. TimeML: Robust Specification of Event and Temporal Expressions in Text. *New directions in question answering*, 3:28–34.

James Pustejovsky, Patrick Hanks, Roser Sauri, Andrew See, Robert Gaizauskas, Andrea Setzer, Dragomir Radev, Beth Sundheim, David Day, Lisa Ferro, et al. 2003b. The Timebank Corpus. In *Corpus linguistics*, volume 2003, page 40.

James Pustejovsky, Bob Ingria, Roser Sauri, Jose Castano, Jessica Littman, Rob Gaizauskas, Andrea Setzer, Graham Katz, and Inderjeet Mani. 2005. The specification language TimeML. *The language of time: A reader*, pages 545–557.

Matthew Richardson and Pedro Domingos. 2006. Markov Logic Networks. *Machine learning*, 62(1-2):107–136.

Gerard Salton, Anita Wong, and Chung-Shu Yang. 1975. A Vector Space Model for Automatic Indexing. *Communications of the ACM*, 18(11):613–620.

Roser Saurí, Robert Knippen, Marc Verhagen, and James Pustejovsky. 2005. Evita: a robust event recognizer for qa systems. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 700–707. Association for Computational Linguistics.

Andrea Setzer, Robert Gaizauskas, and Mark Hepple. 2003. Using Semantic Inferences for Temporal Annotation Comparison. In *Proceedings of the fourth international workshop on inference in computational semantics (ICOS-4)*.

H Gregory Silber and Kathleen F McCoy. 2002. Efficiently Computed Lexical Chains as an Intermediate Representation for Automatic Text Summarization. *Computational Linguistics*, 28(4):487–496.

Josef Steinberger and Karel Ježek. 2009. Update Summarization Based on Novel Topic Distribution. In *Proceedings of the 9th ACM symposium on Document engineering*, pages 205–213. Association for Computing Machinery.

Josef Steinberger, Mijail Kabadjov, Ralf Steinberger, Hristo Tanev, Marco Turchi, and Vanni Zavarella. 2011. JRCS participation at TAC 2011: Guided and multilingual summarization tasks. In *Proceedings of the Text Analysis Conference (TAC)*.

Josef Steinberger. 2007. *Text summarization within the LSA framework*. Ph.D. thesis, University of West Bohemia.

Nicola Stokes. 2004. *Applications of Lexical Cohesion Analysis in the Topic Detection and Tracking Domain*. Ph.D. thesis, National University of Ireland, Dublin.

Hiroya Takamura and Manabu Okumura. 2009. Text Summarization Model Based on Maximum Coverage Problem and its Variant. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 781–789. Association for Computational Linguistics.

Yee W. Teh, Michael I. Jordan, Matthew J. Beal, and David M. Blei. 2004. Hierarchical Dirichlet Processes. *Journal of the American Statistical Association*, 101.

Naushad UzZaman, Hector Llorens, Leon Derczynski, Marc Verhagen, James Allen, and James Pustejovsky. 2013. SemEval-2013 Task 1: TempEval-3: Evaluating time expressions, events, and temporal relations. In *Proceedings of the 7th International Workshop on Semantic Evaluation (SemEval 2013), in conjunction with the Second Joint Conference on Lexical and Computational Semantcis (* SEM 2013). Association for Computational Linguistics, June*.

Marc Verhagen, Robert Gaizauskas, Frank Schilder, Mark Hepple, Graham Katz, and James Pustejovsky. 2007. Semeval-2007 task 15: TempEval temporal relation identification. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, pages 75–80. Association for Computational Linguistics.

Marc Verhagen, Roser Sauri, Tommaso Caselli, and James Pustejovsky. 2010. SemEval-2010 task 13: TempEval-2. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 57–62. Association for Computational Linguistics.

Xiaojun Wan, Jianwu Yang, and Jianguo Xiao. 2007. Manifold-Ranking Based Topic-Focused Multi-Document Summarization. In *IJCAI*, volume 7, pages 2903–2908.

Lu Wang, Hema Raghavan, Vittorio Castelli, Radu Florian, and Claire Cardie. 2013. A Sentence Compression Based Framework to Query-Focused Multi-Document Summarization. In *Proceedings of Association for Computational Linguistics*, pages 1384–1394.

Li Wenjie, Wei Furu, Lu Qin, and He Yanxiang. 2008. PNR: Ranking Sentences with Positive and Negative Reinforcement for Query-oriented Update Summarization. In *Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1*, pages 489–496. Association for Computational Linguistics.

Katsumasa Yoshikawa, Sebastian Riedel, Masayuki Asahara, and Yuji Matsumoto. 2009. Jointly identifying temporal relations with markov logic. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-Volume 1*, pages 405–413. Association for Computational Linguistics.

Renxian Zhang, Wenjie Li, and Qin Lu. 2010. Sentence Ordering with Event-enriched Semantics and Two-layered Clustering for Multi-document News Summarization. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pages 1489–1497. Association for Computational Linguistics.

Dengyong Zhou, Jason Weston, Arthur Gretton, Olivier Bousquet, and Bernhard Schölkopf. 2003. Ranking on Data Manifolds. In *Neural Information Processing Systems (NIPS)*, volume 3, pages 169–176.

Appendix-A

Sample System Generated Summaries

Some System Generated Summaries for Generic MDS System

Following are the example summaries generated by our generic mds system for the document collection of DUC'2004.

(1) Generic Summary for the document set D30001

''Worried that party colleagues still face arrest for their politics, opposition leader Sam Rainsy sought further clarification Friday of security guarantees promised by strongman Hun Sen. Sam Rainsy wrote in a letter to King Norodom Sihanouk that he was eager to attend the first session of the new National Assembly on Nov. 25, but complained that Hun Sen's assurances were not strong enough to ease concerns his party members may be arrested upon their return to Cambodia. Cambodian leader Hun Sen on Friday rejected opposition parties' demands for talks outside the country, accusing them of trying to ''internationalize'' the political crisis. ''

(2) Generic Summary for the document set D30010

''The 18-member Cabinet debate on the Wye River accord would resume only after Yasser Arafat's Palestinian Authority fulfilled all of its commitments under the agreement. A car rigged with explosives blew up Friday morning in Jerusalem's Mahane Yehuda market packed with Israelis shopping for the Jewish Sabbath, killing two people and wounding 21. The radical group Islamic Jihad claimed responsibility Saturday for the suicide bombing of a crowded Jerusalem market and promised more attacks to try to block the new peace accord. Israel should not use Friday's suicide bombing as an excuse to stop the peace process.''

(3) Generic Summary for the document set D30056

''The severed heads of the hostages, three Britons and a New Zealander, were found. Many of the victims are children, who dig the ground to look for ammunition left over from Chechnya's two-year independence war against Moscow. Chechen police were searching for the bodies of four kidnapped foreigners who were beheaded during a botched attempt to free them. Assailants have abducted Chechnya's top prosecutor, who was investigating the killings of four kidnapped foreigners. Cochetel was freed about two minutes into the predawn operation on the border between the republics of Chechnya and Ingushetia.''

Some System Generated Summaries for Query-focused MDS System

Following are the example summaries generated by our query-focused mds system for the document collection of DUC'2007.

(1) Summary for the topic description (document set D0703A) : **"Describe steps taken and worldwide reaction to introduction of the Euro on January 1, 1999. Include predictions and expectations reported in the press."**

''The Euro will be launched on January 1, 1999. Local press reports today said that the building societies have produced blueprints for euro savings accounts, passbooks and cash-dispensers despite the government's increasingly Eurosceptical tone. Despite skepticism about the actual realization of a single European currency as scheduled on January 1, 1999, preparations for the design of the Euro note have already begun. The European single currency euro will go ahead on schedule on January 1, 1999 with a broad membership, according to a survey of some prominent British economists. The report also said Duisenberg expects the future relationship between the dollar and the euro, which officially goes into effect on Jan. 12, to be stable. Stressing that the introduction of a single currency will be a great contribution to the unity of an expanded European Union (EU), Juppe reiterated France's commitment to the timetable and criteria of the single currency system set in the Maastricht treaty, under which the single European currency, recently named Euro, will be realized by January 1, 1999. The Frankfurt-based body said in its annual report released today that it has decided on two themes for the new currency: history of European civilization and abstract or concrete paintings. Europe's new currency, the euro, will rival the U.S. dollar as an international currency over the long term, Der Spiegel magazine reported Sunday. The European Union member states are required to completely replace their own national currencies with the Euro from January 1, 2002. ''

(2) Summary for the topic description (document set D0708B) : **"What countries are having chronic potable water shortages and why?"**

''The move came against a backdrop of a severe water shortage in the country. China is one of the many countries in the world facing water shortage, a situation plaguing more than 300 of its 660-odd cities. Although Nepal is rich in water resources, water shortage is pervasive in the country because of its inability to tap the resources and the lack of well-managed supply system. A Lebanese official has warned that his country is suffering an annual water shortage of more than 1 billion cubic meters, the Daily Star reported Friday. Due to the current drought in the Horn of Africa, water

shortage has reigned throughout the east African country and more than 2.2 million Kenyans are threatened by starvation. Fernandes was speaking to the press on the water shortage problem in Luanda. The water shortage has caused some 6,000 people in the province to move, the report added. In northwestern China, which has half of the country's land, arable land has become increasingly desertified and sandstorms have become more frequent because of improper use of water resources. It is widely believed here that water shortage would be eased to a great extent once these plants become fully operational. It was reported that water shortage brought by the El Nino weather El Nino will be very serious in the Philippines. The Addis Ababa Regional Water and Sewerage Authority announced that the shortage of potable water in the capital city of Ethiopia will be solved in the last quarter of this year.''

Some System Generated Summaries for Guided Update MDS System

Following are the example summaries generated by our guided update mds system for the document collection of TAC'2011.

(1) Summary for the update document set D1109B-B

''Israeli Prime Minister Ehud Olmert pledged Monday that Israel "will not relent" in its struggle against militants, after a suicide bombing in the southern town of Dimona earlier in the day. A violent offshoot of Palestinian President Mahmoud Abbas' Fatah movement claimed responsibility Monday for the suicide attack in the southern Israeli town of Dimona. Israel vowed on Monday to "continue to fight terrorism by all necessary means" after the first suicide blast in a year killed a woman in the desert town of Dimona. The Palestinian Authority condemns the Israeli operation in Qabatiya during which two citizens were killed by the Israeli army.''

(2) Summary for the update document set D1131F-B

''A researcher at Academia Sinica warned that the mushroom coral along the shores of Green Island, off Taitung in eastern Taiwan, is dying due to damage caused by human activities, and he called for action to save the endangered reef. Taiwan's coral reefs are relatively healthier than those in other parts of the world, leading U.S. researchers to team up this week with the island's scientists to study the natural undersea formations, in hopes of saving coral reefs worldwide. Taiwan's coral reefs were under severe stress or had been heavily damaged, and that a trend of declining coral cover deserves attention.''

(3) Summary for the update document set D1143H-B

''Disgraced former American football star Simpson was sentenced to 15 years in prison for armed robbery and kidnapping during a 2007 raid on a Las Vegas hotel room. O.J. Simpson was found guilty of robbery and kidnapping here Friday, 13 years to the day after the American football legend was acquitted of brutally murdering his ex-wife and her friend. O.J. Simpson is in custody in Florida and will be brought before a judge next week on allegations that he violated terms of his release on bail in a Las Vegas armed robbery case. A judge agreed to delay former football star O.J. Simpson's trial on armed robbery and kidnapping charges until September.''