An Information Loss based Framework for Document Summarization

Thesis submitted in partial fulfillment

of the requirements for the degree of

Master of Science (by Research)

in

Computer Science

by

Chandan Kumar

200607003

chandan_kumar@research.iiit.ac.in

International Institute of Information Technology

Hyderabad, India

June 2009

INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY

Hyderabad, India

CERTIFICATE

It is certified that the work contained in this thesis, titled “An Information Loss based Framework for Document Summarization” by Chandan Kumar, has been carried out under my supervision and is not submitted elsewhere for a degree.

Date

Advisor: Prof. Vasudeva varma

The talent of success is nothing more than doing what you can do well.

– Henry Wadsworth

To my Grandpa and Grandma

Acknowledgments

First and foremost I would like to thank my advisor, Dr. Vasudeva Varma for all his guidance, support, encouragement and patience all through my research work. I am grateful to him for providing me an opportunity to work in SIEL along with an open and friendly research environment, giving me an exposure to current research problems and involving me with industry projects and presentations. This has had a great impact on building my research aptitude and skills.

I am thankful to Prasad Pingali for his in depth discussions with me on my research work and providing his valuable feedback and suggestions. I would like to thank Dr. Kannan Srinathan for his motivation and advice in the initial phase of my graduate studies.

I had a great experience in IIIT for the past three years with lot of cherishable memorable moments and I would like to thank all those individuals who have helped me directly and indirectly in carrying out this work. I need to thank Satyanarayan Patel and Narsimha Raju who have been very supportive and helpful to me all the times. I must thank my friends Abu and Prashanth for helping me with writing skills, and for their suggestions and constant moral support. Apart from these I would also like to thank all my labmates at SIEL, batchmates and friends for making my stay at IIIT a memorable one!

Abstract

With vast amount of data available electronically, there has been an ever increasing need to readily sort out and extract only the chief and pertinent sections from the data sources to present to the user. Time and space being critical constraints in this electronic era, it is indispensable to provide a mechanism to locate and browse required information quickly to avoid information overload. Automatic multi-document text summarization fulfills this need by presenting user with desired information in the form of a quick summary. Given a collection of documents related to same topic, the goal of multi document summarization system is to generate a short and concise summary that can be read in lieu of the original document collection. In multi-document summarization, sentence extraction is a critical phase in the formation of useful summaries. where the task is to pick a subset of sentences from the document cluster to form the summary giving an overall sense of the document's content. Developing a principled sentence extraction mechanism that also performs empirically well is a big challenge in itself.

This thesis presents a new sentence extractive summarization framework based on Information Loss. We treat summarization as a decision making problem. Given a set of documents, either to a human or system, the selection of few sentences as a representative to whole document set is a critical decision making problem. As per decision theory we derive a general extraction mechanism for picking sentences based on an ascending order of the expected risk of information loss. We propose to use an intrinsic loss function to compute this information loss and to make a decision on picking a sentence as a part of summary. Sentences and documents are considered as different text units and are represented by probabilistic language models to estimate their distributions. In this inferential setting we use entropy based intrinsic loss function (relative entropy) to measure the discrepancy between the sentence and document model. Relative entropy loss function measures how bad a sentence distribution is in modeling the documents distribution. By doing this

we are able to capture the amount of information loss in picking a sentence to represent the whole document cluster.

The selection of a sentence in summary is not by a set of features but only the measure of loss. A large document set is divided into smaller text units (sentences), each of which tries to approximate the document set variationally, yielding an overall variational approximation. Sentences which are the candidates of summary act as a surrogate for document set in a larger inference process. This decomposition strategy leads us directly to a new sentence extraction algorithm. With a simple redundancy identification and text reformulation mechanism, we come up with a lightweight summarizer to generate more informative summaries. The proposed algorithm generates the extracts on the fly without extensive computation or training which seem to be used in various state of the art algorithms. Furthermore, we consider different information theoretic divergence measures and loss functions to estimate loss between sentence and document distribution, and analyze their performance.

In order to evaluate the performance of our approach, we have used DUC (Document Understanding Conference) and MSE (Multi-Lingual Summarization Evaluation) dataset that have been widely used in recent document summarization evaluations. We have applied ROUGE (Recall-Oriented Understudy for Gisting Evaluation) as the automatic summary evaluation metric which is the standard way of evaluation of summaries. It essentially calculates n-gram overlaps between automatically generated summaries and previously-written human summaries. A high level of overlap indicates a high level of shared concepts between the two summaries. Our overall results are the best reported on the DUC-2004 summarization task for all three metrics ROUGE-1 ROUGE-2 and ROUGE-SU4, and are the best, but not statistically significantly different from the best system in MSE-2005. Results on DUC-2007 dataset supports our results on DUC-2004 and MSE-2005. Our system is also substantially simpler than the previous best systems.

Furthermore, we go beyond the traditional notion of generic relevance and incorporate a user factor as sentence extraction criteria. Here we treat summarization process as not only a function of the input text but also of its reader. We believe that a good summary should change in accordance to preferences of its reader. For this purpose we model the user in the proposed information loss based framework to extract user specific personalized summaries by creating web based profiles using the personal data available online. To evaluate personalized summaries, a controlled user-centered qualitative evaluation was carried out on news articles of science and technology domain. The results indicate better user satisfaction with personalized summaries compared to generic summaries.

Contents

1 Introduction **1**

1.1 Motivation 1

1.2 Problem Description 4

1.2.1 Problem Statement 4

1.3 Solution Outline 5

1.4 Contributions 8

1.5 Organization of the Thesis 9

2 Summmarization Background and Related work **10**

2.1 Automatic Document Summarization 10

2.1.1 Extract vs. Abstract 11

2.1.2 Query Focused vs. Generic Summaries 12

2.1.3 Single Document vs. Multi Document Summaries 12

2.2 How Summaries Help: A Use Case 13

2.3 Major Approaches to Multi Document Summarization 14

2.3.1 Centroids based Approach 14

2.3.2 Supervised Approaches 16

2.3.3 Graph based Approach 16

2.3.4 HMM based approach 17

2.4 Evaluation 18

2.4.1 Intrinsic Evaluation . . . 18
2.4.2 Extrinsic Evaluation . . . 19
2.5 Summary . . . 20

3 An Information Loss Framework for Document Summarization 21
3.1 Decision Theory . . . 21
3.1.1 Loss Function . . . 23
3.2 Probability Estmation of Text . . . 25
3.3 Summarization as Decision Problem . . . 26
3.4 Sentence Selection based on Information Loss . . . 28
3.4.1 Document and Sentence Representation . . . 29
3.4.2 Sentence Expected Risk . . . 30
3.4.3 Different Divergence Measures and Loss Function . . . 32
3.5 Summary Generation . . . 34
3.5.1 Redundancy Identification . . . 34
3.5.2 Sentence Organization . . . 35
3.6 Summary . . . 37

4 Experiments and Evaluation 39
4.1 Evaluation . . . 39
4.2 Evaluation Metric . . . 39
4.3 Dataset . . . 41
4.3.1 DUC 2004 . . . 41
4.3.2 MSE 2005 . . . 44
4.3.3 DUC 2007 . . . 44
4.4 Evaluating Different Divergence Measures as Loss Function . . . 48
4.5 Redundancy Threshold . . . 52
4.6 Effect Of Smoothing . . . 53

4.7 Related work comparision 57
4.8 Discussion 60
4.9 Summary 61

5 Personalization of Summaries: A New Direction **63**
5.1 Motivation 63
5.2 User Specific Summarization 64
5.2.1 User Dependent Loss 65
5.2.2 Document and Sentence Representation 66
5.2.3 Risk Estimation 66
5.2.4 Estimating User Background model: Experimental Setup 67
5.3 Summary Generation 68
5.4 Evaluation 68
5.5 Summary 71

6 Conclusion and Future Work **73**
6.1 Future Directions 75

List of Figures

1.1 Sentence selection for a User 5
1.2 Information Loss when converting document to summary 6
1.3 Information Required when converting summary to document 7

4.1 Summarization performance (ROUGE-2) with respect to different values of redundancy threshold 52
4.2 Summarization performance (ROUGE-SU4) with respect to different values of redundancy threshold 53
4.3 Summarization performance (ROUGE-2) with respect to different values of Jelinek-Mercer smoothing parameter 54
4.4 Summarization performance (ROUGE-SU4) with respect to different values of Jelinek-Mercer smoothing parameter 55
4.5 Summarization performance (ROUGE-2) with respect to different values of Dirichlet prior smoothing parameter 56
4.6 Summarization performance (ROUGE-SU4) with respect to different values of Dirichlet prior smoothing parameter 57

5.1 Average Scores for different Users 70
5.2 Score of Different topics for a User 71

List of Tables

4.1 ROUGE-1 systems comparison on DUC 2004. 42
4.2 ROUGE-2 systems comparison on DUC 2004 43
4.3 ROUGE-SU4 systems comparison on DUC 2004. 43
4.4 Example: System generated and Human written summaries from DUC 2004 dataset 45
4.5 Comparison on MSE 2005 46
4.6 Example: System generated and Human written summaries from MSE 2005 dataset 47
4.7 Comparison on DUC 2007 48
4.8 Example: System generated and Human written summaries from DUC 2007 dataset 49
4.9 Different Loss Functions performance on DUC 2004 50
4.10 Different Loss Functions performance on MSE 2005 51
4.11 Different Loss Functions performance on DUC 2007 51

5.1 Example: System Generated Generic and Personalize Summaries 69

Chapter 1

Introduction

1.1 Motivation

In the recent past there has been an explosive growth in the volume of textual information available electronically. With time being a critical constraint for any user, it has become very important to readily sort out and present the vast amount of data available on web and database archives. This therefore necessitates exploring methods of allowing users to search, locate and browse the required information readily from a collection of documents to avoid information overload. Automatic multi-document text summarization fulfills such an information seeking need by providing a mechanism for the user to quickly view the chief and pertinent sections from a set of documents. The goal here is to take a collection of documents related to same topic, generate a short and concise summary that can be read in lieu of the original document collection.

There are two kinds of approaches to document summarization: abstraction and extraction. Even though efforts have been put to generate an abstract summary, extraction is still the most feasible approach, and most of recent works in this area are based on extraction. Extraction is the process of selecting important units from the original document and concatenating them into a shorter form as summary [37, 7, 13]. In contrast, an abstract may or may not contain words in common with the document [6, 41, 39]. Extractive approach to summarization can employ

various levels of granularity, e.g., keyword, sentence, or paragraph. With readability of a list of keywords being typically low and paragraphs unlikely to cover the information content under space constraints, sentences have emerged to be the most popular text unit for summary generation. So the main challenge is to identify which sentences from the input documents should be included in a summary. Even systems that go beyond extraction (abstraction) and use generation techniques to reformulate or simplify the text of the original articles need to decide which simplified sentences should be chosen, or which sentence should be fused together or rewritten.

Most of the sentence extraction methods till date use various heuristics, linguistic and statistical features to estimate sentence importance [8]. Generally a summarization system uses combination of many such features to extract sentences for summary generation. The presence of many features makes the system very complex to be used in real time systems. Also estimation and parameterization of these multiple features makes the summarization process data dependent, as most of the heuristic and linguistic features differ for different genre of text. Heuristic features like Title-keyword(Sentences containing words that appear in the title are also indicative of the theme of the document), Location heuristic(Position of the sentence), Indicative phrases (Sentences containing key phrases like "this report.....") are not effective for summarization in every case because some features depend on the particular format and the writing style of documents [17, 34]. Linguistic features have some difficulties in requiring to use high quality linguistic analysis tools such as discourse parser and linguistic resources such as WordNet, Lexical Chain, and Context Vector Space [39, 8]. They are useful but require much memory and processor capacity because of additional linguistic knowledge and complex linguistic processing. Statistical features (topic signature based, cluster based, centroid based etc.) based on several text mining techniques to compute sentence importance are very common. But generally they have been used in combination with other features because they didn't show great performance as an individual feature. Supervised approaches have also been tried extensively, where the sentence classifiers are trained using human-generated summaries as training examples for feature extraction and parameter estimation [34, 35, 47]. The major drawbacks of the supervised approaches are domain dependency and the problems caused

by the inconsistency of human generated summaries. So the use of multiple features or extra information sources for training makes the summarization system inconvenient to be used in real time applications.

There is a need of effective extraction mechanism which can produce sentence extract on the fly without extensive computation or training or the estimation and parameterization of multiple features. This motivates us to derive a new theoretical mechanism to formalize the task of picking sentence towards summary, which do not use several empirical or heuristic features or trained classifiers and can achieve good performance.

The main challenge of document summarization system is to decide which sentences from the input documents should be included in a summary and there is a great amount of uncertainty involved in the process. Statistical decision theory provides a theoretical foundation to deal with problems of action and inference under uncertainty and has been used successfully in various text based applications. This motivates us to treat sentence extractive summarization task as decision making problem. We try to come up with a novel extraction mechanism for document summarization motivated from statistical decision theory.

As a follow up to generic summarization, this thesis also explores the possibility of incorporating user specific content in summary generation. Most of the current summarization systems produce one uniform summary for all users without considering the user's personal interest. But different users may have different perspectives on the same text, based on their field of expertise and interest. It is clear through various experiments in summarization literature that when persons of different background and expertise summarize the same articles, they include different content from each other, reflecting their personal interest and background knowledge. Thus there is a great need for summaries to cater to the user's interests. So a good summary should change in accordance to preferences of its reader. So this motivates us to explore the possibility of adding user factor into automatic summarization process and coming up with a design to generate personalize summaries for end user.

1.2 Problem Description

In this thesis we aim to reduce the problem of information overload by providing a multi document summary to the end user. The goal here is to take a collection of documents related to same topic and generate a short and concise summary that can be read in lieu of the original document collection. So user need not go through each document and can save his time and efforts.

In multi-document summarization, sentence extraction is a critical phase in the formation of useful summaries. In this thesis we address this challenging task of developing a sentence extraction mechanism. The task here is "To extract content from an information source, that is a collection of related documents, while removing redundancy and taking into account similarities and differences in the information content, and present the most important content to the user". More specifically system needs the ability to find and extract most important sentences across the documents. Since the documents may be authored by different persons and are about a related event, there will be higher level of redundant information across the documents, this demands for an anti-redundancy measures. Finally the system should have the ability to combine extracted sentences in a useful and readable manner.

We also address the problem of personalized summarization, i.e. from the same text, generation of different summaries for different users based on their interest and background knowledge. As the main phase of summary generation process is sentence extraction, the problem of personalized summary generation can be decomposed to user specific sentences extraction. So in the later part of this thesis we concentrate on the problem of "how to model user in a sentence extraction process to generate personalized summaries."

1.2.1 Problem Statement

In this Thesis, we address the problem of multi-document summarization through a sentence extractive procedure. Here the task is to pick subset of sentences from the document cluster (set of documents to be summarized) and present them to user in form of summary that provides an over-

all sense of the document's content. As an extension we also try to come up with a user specific sentence extraction procedure to generate personalized summary.

1.3 Solution Outline

Automatic Summarization is defined as a process whose goal is to produce a condensed representation of the content of its input for human consumption. In automatic document summarization, input is documents which are viewed as information sources. So we can view summarization as a condensation or compression process of an information source. We know the compression or condensation process of any information source suffers from loss of information. In this thesis we try to model this information loss to setup a novel sentence extraction mechanism for text summarization.

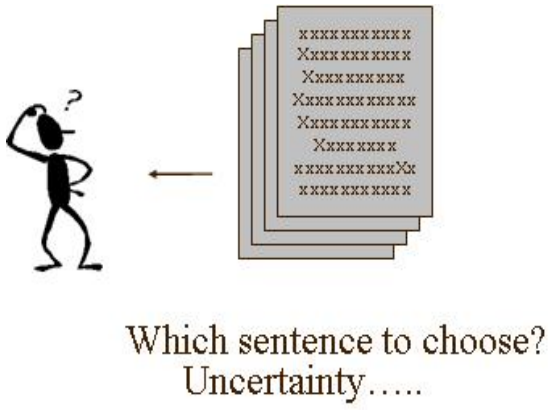

Figure 1.1: Sentence selection for a User

We take the scenario of human doing sentence extraction task. Given a document or set of documents to summarize, to a human, the selection of few sentences as a representative to whole document set(which contains hundreds of sentences) is a tough task[fig 1.1]. There is a great amount of uncertainty involved about which sentence to choose as part of summary. This process can be viewed as a critical decision making problem. As per decision theory there is a risk involved

in taking each decision [10, 11]. So we treat summarization as a decision making problem where there is a risk involved in the selection of sentences to generate summary. The risk involved here is the measure of information loss when we pick a particular sentence instead of whole document set[fig 1.2]. To understand the risk factor in sentence extraction we can take a real world example. When a financial executive has to make a critical decision, she analyzes the risk of economic loss. In the same manner a document summarization system should analyze the risk of information loss, as the documents are all about information.

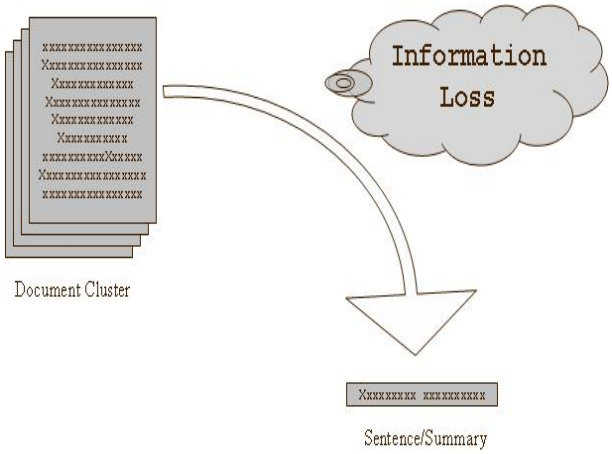

Figure 1.2: Information Loss when converting document to summary

To compute the information loss associated with a particular sentence with respect to the whole document cluster we propose to make use of loss function between sentence and document. As the context is inferential, we use the loss function to measure the discrepancy between the models of sentence and document. We define sentence and documents as two difference probabilistic models using unigram language model. We use the entropy based intrinsic loss function [4] (relational entropy) to compare sentence with document model. Relative entropy loss function measures how bad a sentence distribution is in modeling the documents distribution. This information theoretic measure actually estimates the amount of information required to reconstruct the original source document set, which is equal to the amount of information loss when we pick that particular sen-

tence in summary to represent the content of document set[fig 1.3]. Sentence with minimum information loss are most important to be included in summary. The idea is to view summarization as method to approximate a large document set by a smaller summary with minimum information loss. In this view, selection of a sentence in summary is not by a set of features (heuristic and experimental used in previous approaches), but only the measure of loss. We try to approximate a document set by a small sentence set, or approximating large text with small text. The candidate of summary act as a surrogate for document set in a larger inference process.

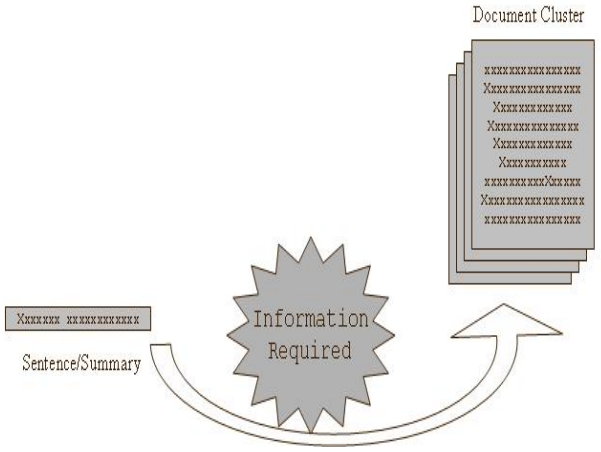

Figure 1.3: Information Required when converting summary to document

We have evaluated performance of our approach on DUC (Document Understanding Conference) and MSE (Multi-Lingual Summarization Evaluation) dataset that have been widely used in recent document summarization evaluations. Our approach performs better as compared to top performing systems. We are able to achieve good results using an approach that does not require any extensive computation of topic themes. Neither does it need any external knowledge source other than document cluster, nor the estimation and parameterization of different features.

We extended the proposed approach further to generate personalized summaries. We make use of user's personal data available on web to model their profiles. We use this user profile data to influence the document model as each term in document set may have different interest for different users. So we estimated user specific model to incorporate user preference in the loss estimation

process for sentence extraction. A controlled user-centered qualitative evaluation was carried out on news articles of science and technology domain. The results indicate better user satisfaction with personalized summaries compared to generic summaries.

1.4 Contributions

- **Summarization: An Information loss Problem**

  This thesis gives a decision theoretic perspective to document summarization and formalize sentence extraction summarization as information loss problem. Given a document or set of documents to summarize, either to a human or system, the selection of few sentences as a representative to whole document set(which contains hundreds of sentence) is a critical decision making problem. As per Bayesian decision theory we define sentence selection in terms of risk of information loss, and sentences with minimum expected risk will be chosen as part of summary.

- **A light-weight and Effective Algorithm for summary Generation**

  Through this formulation of information loss we come up with a very light-weight function to generate more informative summary than the earlier approaches which use very complex algorithm for summary generation. The proposed extraction mechanism generates sentence extract on the fly without extensive computation or training or the estimation and parameterization of multiple features which seem to be used in various state of the art algorithms. We know that summarization is information access technology which has always suffered from computational complexity that makes it inconvenient to use in real-time applications. With the proposed framework we can build summarization applications that are real-time and does not require too much of computational power.

- **Personalization of Summaries.**

  Current summarization systems produce a uniform version of summary for all users. However summaries which are generic in nature do not cater to the user's background and in-

terests. We propose to make the summarization process user specific and present a design for generating personalized summaries of online articles that are tailored to each person's interest. We have designed an experimental setup where we propose a web based profile creation system for knowledge workers using their personal data available on web to model their profiles. In the university environments this is an effective way to carry out evaluation and experiments related to user modeling.

- **Study the effect of Divergence measures, Redundancy and Smoothing on summarization performance.**

  We analyze the performance of different divergence measures as loss functions in the sentence extraction framework. We also studied the effect of different values of redundancy threshold on summarization performance and found that a 40 percent term overlap between sentences is a good heuristic to estimate redundancy. Studying the effect of smoothing probabilities of language models is another contribution of this thesis. We evaluated popular smoothing methods (Jelinek-Mercer, Dirichlet priors), and found that the summarization performance is not that sensitive to the smoothing parameters as compare to its impact on information retrieval systems.

1.5 Organization of the Thesis

In chapter 2 we discuss the major classification of text summarization in order to explain what the task is, and what it involves. We also discuss the major approaches of multi document summarization to explain what has been achieved in previous research. Next the theory behind the decision problem and loss functions is set out in chapter 3, and we propose our information loss based sentence extraction framework for document summarization. In chapter 4, we give the experimental evaluation and results. In chapter 5 we discuss our idea of incorporating user in the summarization process and presented a design to generate personalized summary. In chapter 6 we conclude our work with the set of possible future directions.

Chapter 2

Summmarization Background and Related work

2.1 Automatic Document Summarization

It is hard to imagine everyday life without some form of summarization, A preview or trailer of show is a summary. Abstract of scientific articles are summary written by authors, a table showing cricketer statistic for a season is very much a summary. Other examples are reviews (books and movies), minutes of a meeting, a program for conference, a weather forecast, a resume, a stock market bulletin, a library catalog, abstracts of articles in news journals, a web page listing resource in a particular subject area, a catalog of various products available from a vendor can be a summary.

In this thesis our focus is on document summarization i.e. summarization of documents which are viewed as information sources whose contents reflects things in the world. When the document summarization is carried out by machines, it is termed as Automatic Document Summarization. Though the automated summarization dates back to Luhn's work at IBM in 1950's [6], the research and development in the area has grown in importance in the late twentieth century with the rapid growth of the World Wide Web. With time being a critical constraint, it becomes necessary to access most relevant information quickly. As Inderjeet Mani puts it, goal of Automatic Text

Summarization is “To extract content from an information source and present the most important content to the user in a condensed form and in a manner sensitive to the user’s or an application’s need [8]”.

Summaries can be classifed into many different categories , Here we will discuss some major categories:

2.1.1 Extract vs. Abstract

Summaries that are constructed by extracting important passages, sentences or phrases from the source document are called extracts. In contrast, an abstract may or may not contain words in common with the document [41, 39, 9]. Authors using abstraction techniques are not constrained as those using extractive ones, and can summarize a wider range of materials effectively and often with smaller amount of text.

While there has been some efforts to generate abstract summaries that requires using heavy machinery from natural language processing, including grammars and lexicons for parsing and generation [15], Extraction is still the most feasible approach, and most of recent works in this area are based on extraction [32, 31, 38]. Extraction is the process of selecting important units from the original document and concatenating them into a shorter form as summary. We also concentrate on the shallow approach of document summarization based on extraction.

Why Sentence Extraction: Extractive approach to summarization can employ various levels of granularity, e.g., keyword, sentence, or paragraph. We concentrates on sentence-extraction because the readability of a list of keywords is typically low while paragraphs are unlikely to cover the information content of a document given summary space constraints. There is also a linguistic motivation for preferring sentence extraction rather than paragraphs. The paragraph is not a traditional linguistic unit in linguistic literature, and being specific to written text, it reflects publishing and formatting conventions. The sentence on the other hand has historically served as a prominent unit in syntactic and semantic analysis (while paragraph and documents of various kinds are not). In particular, logical accounts of meaning offer precise notions of sentential meaning. Such

notions can be extended to discourses as a whole; though of course they begin with a sentential representation.

2.1.2 Query Focused vs. Generic Summaries

Automatic text summarization systems often produce generic summaries that highlight the most salient points of a given text. However query focused summarization system has access to the query/question and summary should adapt to suit the query given [31].

A query-focused summary presents the content that resembles the query. It is essentially a process of retrieving the query relevant sentences/passages from the document. Most of the query focused summarization systems is based on conventional IR technologies, as the objective here is quite similar to text retrieval process. On the other hand, a generic summary provides an overall sense of the document's content, aimed to a broad readership community and reflects the fundamental aim of automatic summarization process [37]. Also there is neither query nor topic provided to the generic summarization process; it becomes even more challenging to develop a high quality generic summarization method. In this thesis we focus on generic summarization process.

2.1.3 Single Document vs. Multi Document Summaries

To generate a single output that summarizes the salient points across multiple documents is more difficult. Since the documents are related by a common topic, they likely contain similar content; thus a system cannot simply concatenate many single document summaries together.

The trend of text summarization has moved from single document summarization to more complex and challenging problem of summarizing multiple documents [30]. Here the goal is to produce a summary that can be used to concisely describe the information contained in a cluster of documents and facilitate users to understand the document cluster. In this thesis our focus is on more complex and challenging problem of summarizing multiple documents.

In this Thesis, we address the problem of generic multi-document summarization through a

sentence extractive procedure. Here the task is to pick subset of sentences from the document cluster (set of documents to be summarized) and present them to user in form of summary that provides an overall sense of the document's content.

2.2 How Summaries Help: A Use Case

The topic of this thesis is generic multi-document summarization. It is thus appropriate to show a use case i.e. how real users can benefit from such summaries. Here we briefly discuss a task based evaluation [61] showing that generic summaries help users writing a report to produce better reports and to make their experience with news browsing system better. The automatic summaries produced by Newsblaster [59, 60] were used for evaluation. Newsblaster is a system that provides an interface to browse the news, featuring multi-document summaries of clusters of article related to same event. The users were also exposed to alternative interfaces: one featuring no summaries at all, and one featuring a first sentence summary for each document, as well as a one sentence centroid summary for each cluster. Human subjects used the different interfaces with no access to other information but provided through the interface. They were asked to write reports answering three questions for each of three events. For each event, there were 4 clusters of about 10 news articles: two directly related to the topic and two peripherally related. The human subjects were asked to collect facts that answer the questions they were given, and after completing each report they were asked about their experience writing the report. There were 13 subjects writing reports using an interface with no summaries, 11 using lead-sentence summaries, and 10 using Newsblaster summaries. The report written using Newsblaster summaries were significantly better than those written using an interface with no summaries. In addition, the subjects felt they had enough time to write the report, they read less of the source documents and felt that the reporting writing task was easier when they had Newsblaster summaries than when no summaries were provided. The full details of the study can be find in [59, 60]. The results of the study confirm that generic summarization is a useful task and that automatically produces summaries can lead to improvements

of reports quality and user satisfaction, motivating the need for summarizer development.

2.3 Major Approaches to Multi Document Summarization

The goal of Multi Document Summarization is "To extract content from an information source that is a collection of related documents and extract content from it, while removing redundancy and taking into account similarities and differences in the information content, and present the most important content to the user in a condensed form and in a manner sensitive to the user's or an application's need [8]." So as per the above definition a multi-document summarization system need the ability to find and extract the main points across the documents considering the cross document references also. Since the documents may be authored by different persons and are about a related event, there will be higher level of redundant information across the documents, this demand for better anti-redundancy measures. The system should have the ability to combine text passages in a useful and readable manner even when the material stems from documents written in different styles. Applications of multi-document summarization are articles generated from various information sources such as online news, blogs, emails, search results. There have been many approaches to produce sentence extractive summaries; here we will discuss some of the more popular mechanisms.

2.3.1 Centroids based Approach

Radev [13] pioneered the use of cluster centroids to play a central role in summarization. He presented a centroid-based MEAD summarization system. The most appealing feature is the fact that it does not make use of any language generation module, unlike most previous systems. All documents are modeled as bags-of-words. The first stage consists of topic detection, whose goal is to group together news articles that describe the same event. To accomplish this task, an agglomerative clustering algorithm is used that operates over the TF-IDF vector representations of the

documents, successively adding documents to clusters and recomputing the centroids according to

$$c_j = \frac{\sum_{d \in C_j} \tilde{d}}{|C_j|} \tag{2.1}$$

where c_j is the centroid of the j-th cluster, C_j is the set of documents that belong to that cluster, its cardinality being $|C_j|$, and $\tilde{d}$ is a truncated version" of d that vanishes on those words whose TF-IDF scores are below a threshold. Centroids can thus be regarded as pseudo-documents that include those words whose TF-IDF scores are above a threshold in the documents that constitute the cluster. Each event cluster is a collection of news articles from multiple sources, chronologically ordered, describing an event as it develops over time. The second stage uses the centroids to identify sentences in each cluster that are central to the topic of the entire cluster. Two metrics are defined, The first accounts for how relevant a particular sentence is to the general topic of the entire cluster; the second is a measure of redundancy among sentences. For each sentence , three different features are used:

- Its centroid value (C_i), defined as the sum of the centroid values of all the words in the sentence,
- A positional value (P_i), that is used to make leading sentences more important. Let C_{max} be the centroid value of the highest ranked sentence in the document. Then $Pi = \frac{n-i+1}{n} C_{max}$.
- The first-sentence overlap (Fi), defined as the inner product between the word occurrence vector of sentence i and that of the first sentence of the document.

The final score of each sentence is a combination of the three scores above minus a redundancy penalty R_s for each sentence that overlaps highly ranked sentences.

Once the documents are clustered, sentence selection from within the cluster to form its summary is local to the documents in the cluster. The IDF value based on the corpus statistics seems counter-intuitive. Both the position factor and the first-sentence similarity factor heavily weight the first few sentences of the documents in the cluster. Thus this metric is genre-specific and applies primarily to newswire articles. For other articles such as technical papers, the scoring would have to be re-designed.

2.3.2 Supervised Approaches

Given a set of training document and their extractive summaries, the summarization process is modeled as a classification problem: sentences are classified as summary sentences and non-summary sentences based on the features that they possess [34, 35]. Features that are used to distinguish summary sentences are Keyword-occurrence(sentences with keywords that are most often used in the document usually represent theme of the document), Title-keyword(Sentences containing words that appear in the title are also indicative of the theme of the document), Location heuristic(In Newswire articles, the first sentence is often the most important sentence; in technical articles, last couple of sentences in the abstract or those from conclusions is informative of the findings in the document), Indicative phrases (Sentences containing key phrases like "this report.....") Short-length cutoff (Short sentences are usually not included in summary). Upper-case word feature (Sentences containing acronyms or proper names are included). The classification probabilities are learnt statistically from the training data, using Bayes' rule:

$$P(s \in < S|F1, F2, ..., FN) = P(F1, F2, ..., FN|s \in S) * P(s \in S)/P(F1, F2, .., FN) \quad (2.2)$$

where s is a sentence from the document collection, $F1, F2..FN$ are features used in classification, S is the to be generated, and $P(s \in < S|F1, F2, ..., FN)$ is the probability that sentence s will be chosen to form the summary given that it possesses features $F1, F2, ...FN$.

2.3.3 Graph based Approach

Some of the most newly developed multi document summarization systems map summarization to graph problems, mainly using PageRank algorithm [23, 21] Of these the most successful application to multi document summarization was that of Erkan and Radev. In their LexRank algorithm, each sentence defines a node in the text graph. To define edges in the graph cosine similarity between two sentences is computed and an edge is added between the nodes representing the two sentences if the similarity exceeds a predetermined threshold. Thus the edges are defined for sentences that share the same words. The PageRank algorithm [18] is then used iteratively to

compute the rank of each sentence as a function of the number of neighbors of each node. The iteration distributes the weight across the graph, and quickly converges to stable node weights. In their discussion of the approach, the authors explain that the iterative spreading of importance in the graph is similar to a voting process: sentences from the entire graph vote for sentences with which they share word overlap. Such a voting procedure can be achieved by a direct frequency count, rater than distributing importance little by little through the nodes. So the pagerank algorithm can be seen as a complex (unobservable) function that assigns weights to sentences based on the frequency of words. To avoid redundancy, sentences of high importance but very similar to more important sentences are not included in summary.

2.3.4 HMM based approach

CLASSY [26, 27] the previous best system on DUC 2004 and MSE 2005 consists of two core components - a Hidden Markov Model for selecting sentences from each document and a pivoted QR algorithm for generating a multi-document summary. The HMM has two kinds of states, which correspond to summary and non-summary sentences in a single document on the basis of some signature terms. The model uses number of signature terms in each sentence as a feature. These terms are decided by the statistic computation similar to [29], derived based on a large set of documents in advance. In addition, the best number of the HMM states needs to be determined based on empirical testing, and the HMM model needs to be learned using training data. After applying the HMM, the top scoring sentences of each document form a weighted token-sentence matrix. A pivoted QR algorithm is then used for scoring and selecting sentences to form the output summary. In addition to these two core components, CLASSY also incorporates a linguistic component as a preprocessing stage to provide the summarization engine simplified sentences as input.

Other than this Hardy etc al. [36] uses passage clustering to detect the topic themes and then extracts sentences which reflect these main themes. Lin and hovy [30] selects important content using sentence position, term frequency, topic signature and term clustering. Yih etc al. [32] uses

machine learning to find content terms using frequency and position information and after that a search algorithm to find the best set of sentences that can maximize the content term scores. Harabagiu and Lacatusu [28] have investigated five different topic representations for extraction. These approaches requires extensive computation of topic themes or signatures, and also they rely on various feature estimation and their parameterization which makes them domain and language dependent.

2.4 Evaluation

Evaluating the quality of a summary has proven to be a difficult problem, principally because there is no obvious ideal summary. Earlier researchers used human evaluation of the summaries which is very expensive to repeat the experiments. Methods for evaluating text summarization (in fact, natural language processing systems) can be broadly classified into two categories [62]. The first is intrinsic evaluation, tests the summarization system itself. The second method is an extrinsic evaluation which tests the summarization system based on how it affects the completion of some other task. Regarding summarization, intrinsic evaluations have assessed mainly coherence and informativeness of the summaries, while extrinsic evaluations have tested the impact of summarization on other tasks like retrieval, text categorization, etc. Intrinsic evaluation is the widely accepted mechanism to evaluate the performance of summarization system, especially for the content evaluation.

2.4.1 Intrinsic Evaluation

Intrinsic evaluation metrics test the summarization task itself, in terms of the coherence and content of the generated summaries. Two broad classes of metrics have been developed: form metrics or coherence and content metrics or informativeness. Form metrics focus on grammaticality, overall text coherence, and organization. These metrics are assessed by humans with grading of the summaries on the basis dangling anaphora, gaps in rhetorical structure etc. [63]. Content or

informativeness of the summary is more difficult to measure. Typically, system output is compared sentence by sentence or fragment by fragment to one or more human-made reference abstracts/extracts, and as in information retrieval, the percentage of extraneous information present in the system's summary (precision) and the percentage of important information omitted from the summary (recall) are recorded. Measures like κ kappa [65] and relative utility [64], both of which take into account the performance of a summarizer that randomly picks passages from the original document to produce an extract, are also used to evaluate the informativeness of the summaries. Document Understanding Conferences (DUC), sponsored by the Advanced Research and Development Activity (ARDA) and run by the National Institute of Standards and Technology (NIST), helps researchers to further progress in summarization by enabling researchers to participate in large-scale experiments. In DUC-01 and DUC-02 competitions NIST used the Summary Evaluation Environment (SEE) interface to support summarization evaluation. In DUC-04 the summaries are evaluated using ROUGE metric [7] , an n-gram co-occurrence metric based on machine translation evaluation metric BLEU [66] . It was shown that the ranking of the systems obtained using ROUGE co-related well with the human evaluations.

2.4.2 Extrinsic Evaluation

Summarization can be applied to a variety of tasks including relevance assessment in information retrieval, reading comprehension tasks, text categorization and etc. One can examine the usefulness of the summary with respect to the goal and hence evaluate the summarization system. The Summarization Evaluation Conference (SUMMAC) [16] included three extrinsic evaluation tests: the categorization task (how well can humans categorize a summary compared to its full text), the ad hoc information retrieval task (how well can humans determine whether a full text is relevant to a query just from reading the summary) and the question task (how well can humans answer questions about the main thrust of the source text from reading just the summary). But the interpretation of the results was not simple; studies [64, 67] show how the same summaries receive different scores under different measures or when compared to different ideal summaries created

by humans.

In DUC workshops, evaluation of single and multiple document summarization systems was also done using a question-answering experiment. A group of analysts prepares a set of questions that can be answered by reading important text fragments in a document. Summarization systems automatically produce summaries of the individual documents. Human analysts attempt to answer the questions produced by the analysts, before reading any text/summary (to factor out their background knowledge) and after reading the summary produced by a system and after reading the whole document. The more questions can be answered using an automatically produced summary, the better the system that produced that summary. From the DUC experiments from 2001-2004, it was shown that humans are better summary producers than machines and that, for the news article genre, certain algorithms do in fact do better than the simple baseline of picking the lead material.

2.5 Summary

In this chapter, we discussed major categories of summaries like abstract/extract, query-focused/generic, and single/multiple document summaries. Then we explained the motivation behind addressing the problem of sentence extractive-generic-multi document summarization in this thesis. We then presented a motivational result of a user study, which shows that summarization helps information seeking users in a news browsing site, by helping them to find more relevant information as well as making their experience more pleasant. We then proceeded towards different extractive approaches of document summarization. We also described the functionality of each of these major approaches to extract important sentences. After this, we discussed how the evaluation of the quality of a summary has proven to be a difficult problem in summarization literature. We then described the two basic categories of summarization evaluation viz., intrinsic and extrinsic evaluation. Intrinsic evaluation tests the summarization system itself while the extrinsic evaluation tests the summarization system based on how it affects the completion of other tasks.

Chapter 3

An Information Loss Framework for Document Summarization

In this chapter, we present the Information Loss framework for document summarization. Since the formulation is based on decision theory and probabilstic language modeling, we first give a brief introduction to both.

3.1 Decision Theory

Decision theory, as the name implies, deals with the problem of making decisions. A decision problem in itself is not complicated to comprehend or describe and can be simply summarized with a few basic elements. Decision models lend themselves to a decision making process which involves the consideration of the set of possible actions from which one must choose, the circumstances that prevail and the consequences that result from taking any given action. The optimal decision is the choice which results in most favourable consequences possible.

The uncertainty in decision making, defined as an unknown quantity θ, describing the combination of prevailing circumstances and governing laws, is referred to as the state of nature [2]. In many real problems and those most pertinent to decision theory, the state of nature is not com-

pletely known. Since these situations create ambiguity and uncertainty, the consequences and subsequent results become complicated. Nevertheless, in order to construct a mathematical framework in which to model decision problems, while providing a rational basis for making decisions, a numerical scale is assumed to measure consequences. Usually something is known about the state of nature, allowing a consideration of a set of states as being admissible (or at least theoretically so), and thereby ruling out many that are not. It is sometimes possible to take measurements or conduct experiments in order to gain more information about the state.

A decision process is referred to as statistical when experiments of chance related to the state of nature are performed. The results of such experiments are called data or observations. These provide a basis for the selection of an action defined as a statistical decision rule. Statistical decision theory is concerned with the making of decision in the presence of statistical knowledge which shed light on some of the uncertainties involved in the decision problem. Decision maker should have the knowledge of consequences of the decisions. Often this knowledge quantified by determining the loss that would be incurred for each possible decision and for the various possible values of θ. The incorporation of a loss function into statistical analysis was first studied by Abraham wald 1950 [2]. To summarize, the ingredients of a decision problem include (a) a set of available actions, (b) a set of admissible states of nature, and (c) a loss associated with each combination of a state of nature and action. In statistical decision problem However, observations from an experiment defined by the state of nature are included with (a) to (c).

Lets assume that an observation x is made on a random variable X whose distribution depends on the parameter θ, which is the true state of the nature. Let $A = \{a_1, a_2, .., a_n\}$ be all the possible actions about θ. For given data set x, to choose an action a^* is a decision problem, where the action space is the class A of possible n values. As we discussed above, foundations dictate [10, 58] that to solve this decision problem it is necessary to specify a loss function $L(a, \theta)$ measuring the consequences of acting as if the true value of the quantity of interest were a when the actual parameter value is θ. This loss function $L(a, \theta)$ specifies our decision preferences.

The use of loss functions also been formalized in terms of risk function, In decision theory and

estimation theory, the risk of an action is the expected value of the loss function. The task is to make a decision on which action to take. In order to evaluate each action, we consider the expected loss (or risk) associated with taking action a_i

$$Risk(a_i) = \int L(a_i, \theta)\pi(\theta)d\theta \tag{3.1}$$

The risk based decision making depends on the measure of loss function. For any given model, the risk of an action depends on the loss function $L(a_i, \theta)$ and the prior distribution $\pi(\theta)$. A loss function is a function that maps an event (technically an element of a sample space) onto a real number representing the cost or regret associated with the event. Less technically, a loss function represents the loss associated with an estimate being ”wrong” (different from either a desired or a true value) as a function of a measure of the degree of wrongness.

In decision problem the action space, in principle, consists of all the possible actions that the system can take. An optimal decision is to choose a Bayes action, i.e., an action a^* that minimizes the risk (loss).

$$a^* = arg_n min Risk(a_i) \tag{3.2}$$

3.1.1 Loss Function

As discussed in previous section a loss function represents the loss associated with an estimate (action) being ”wrong” as a function of a measure of the degree of wrongness. Let $L(\delta_e, \delta_a)$ be a loss function measuring the consecuence of estimating δ_a (the actual true parameter value) with δ_e. Conventional loss functions typically depend on the particular metric used to index the model, being defined as a measure of the distance between the parameter and its estimate. They compare δ_e to δ_a. Examples are squared error loss, zero-one loss and absolute error loss.

In a purely inferential context, one should rather be interested in the discrepancy between the models labelled by the true value of the parameter and its estimate [4]. It is suggested that invariant loss functions should be used, more precisely it is argued that, in a purely inferential context, the loss function $L(\delta_e, \delta_a)$ should not be chosen to measure the discrepancy between δ_e and δ_a, but

to directly measure the discrepancy between the models $p(.|\delta_e)$ and $p(.|\delta_a)$ which they label. This type of intrinsic loss is typically invariant under reparametrization, and therefore produces invariant estimators.

So a loss function of the form $L(\delta_e, \delta_a) = L\{p(.|\delta_e), p(.|\delta_a)\}$ is called an **i**ntrinsic loss [4]. In mathematical statistics, intrinsic loss functions are used to measure the distance between statistical models. Intrinsic loss function compares $p(.|\delta_e)$ to $p(.|\delta_e)$. Bernardo [57] and Bernardo and Smith [58] argue that statistical inference is well described as a formal decision problem, where the terminal loss function is a proper scoring rule. One of the most extensively studied of these is the relative entopy(directed logarithmic divergence). The relative entropy function between two models $p(.|\delta_a)$ and $p(.|\delta_e)$ for data $x \in X$ is

$$RE(.|\delta_e, .|\delta_a) = \sum_x p(x|\delta_e) log \frac{p(x|\delta_e)}{p(x|\delta_a)}$$

$$\begin{aligned} RE(\delta_e, \delta_a) &= \sum_x p(x|\delta_e) log p(x|\delta_e) - \sum_x p(x|\delta_e) log p(x|\delta_a) \\ &= \sum_x p(x|\delta_e) \frac{1}{p(x|\delta_a)} - H(\delta_e) \end{aligned} \tag{3.3}$$

The first term $p(x|\delta_e)\frac{1}{p(x|\delta_a)}$ is the cross entropy and the second term $H(\delta_e)$ is the entropy, which is how much we could compress symbols if we know the true distribution.

Relative entropy has an intuitive interpretation, since it is either zero when the probability distributions are identical or has a positive value, quantifying the difference between the distributions. It gives the number of bits which are wasted by encoding events from the distribution $p(x|\delta_a)$ with a code based on distribution $p(x|\delta_e)$. It is a loss function between two probability distributions which measures how bad a given probability distribution is in modeling the other one.

It is a suitable objective function for the overall statistical decision making process. The relative entropy acquires a fundamental information-theoretic interpretation: it corresponds to the mutual information between a random variable representing the choice of the target distribution and the random variable representing a sequence of observations. Hence, it tells us how much information about the target distribution we have gained by our observations. This constitutes an elegant

connection between information theory and statistics. Furthermore, this connection also extends to other fields [5]: In statistical mechanics, the relative entropy can be related to the free energy, that is, the energy that is available for doing thermodynamic work governing the non-equilibrium dynamics; In mathematical finance, it corresponds to the expected reduction in the logarithm of the compounded wealth due to the lack of knowledge of the true distribution; In information theory, the relative entropy [1] can be interpreted as the expected extra codeword-length of an optimal code based on a (wrong) distribution with respect to the optimal code based on the true distribution. This work will emphasize the last interpretation, since it provides a simple and concrete metaphor to reason about relative entropy.

3.2 Probability Estmation of Text

In this thesis we deal with text summarization, for decision theoretic formalization we need to represent the text units in terms of probabilistic models. Probabilistic approaches have been used with great success in many language-oriented tasks [33], including machine translation, part-of-speech tagging, speech recognition, and search engines. The probabilistic approach we are using is commonly known as language modeling. A statistical language model is a probabilistic mechanism for represenation of text [19, 20]. In general a language model is a probability distribution on a finite feature set. These features vary depending on the type of application (words in our case). A language model describes the language generation process and is generally used as a method of generating text or some other desired output. However, in the context of this work, we use language models simply to express the information in a document set. The most common types of models that utilize this assumption is unigram language model, that uses the simplest probability function give by relative counts. For a given piece of text T, we write the probability of the word w, given the model M_T of the text as $P(w|M_T)$.

$$P(w|M_T) = \frac{T(w)}{|T|} \tag{3.4}$$

where $T(w)$ is the count of word w in the text T and $|T| = \sum_w T(w)$ is the text's length.

More complex language models may capture word orders or even the structure of the text [19, 20]. For example, trigram language models would generate a word based on the two previously generated words, thus capturing the local ordering of words. Probabilistic context-free grammars are structure-based generative models of text, allowing for the incorporation of structural constraints. In general, statistical language models provide a principled way to capture quantitatively the uncertainty associated with the use of natural language, and they have found many applications in a variety of language technology tasks, especially speech recognition. Although we ultimately need to explore more sophisticated models, in this thesis, we only use the simplest unigram language model becuase more sophisticated language models would significantly increase the computational complexity for summarization, making it practically infeasible. Also when estimating a document language model or a sentence language model, we are working on an extremely limited amount of data, so the estimation of complex models may not be reliable. In other words, the sparseness of data puts a constraint on the complexity of the model that we can estimate accurately.

3.3 Summarization as Decision Problem

As dicussed in section 3.1 Statistical decision theory provides a theoretical foundation to deal with problems of action and inference under uncertainty. If there are several possible actions, the task is to make a decision on which action to take. In order to evaluate each action, we consider the expected loss associated with taking a particuler action.

Sentence extractive summarizer can be regarded as a decision making system, where, given a document or set of documents, system needs to choose a subset of sentences and present them to the user to convey the information contained in the document set. We consider the sentence

extraction process as a decision making task, where each sentence is a possible action that can be taken, and system has to make a decision on which sentences to pick as part of summary.

Lets consider a document cluster, D, to summarize $D = \{D_1, ..., D_n\}$. Where each document contains a set of sentences $D_i = \{s_1, ..., s_n\}$. For simplicity, we represent the document cluster as the set of all sentences from all the documents present in cluster, i.e., $D = \{s_1, ..., s_k\}$. An extractive summary $S = \{s_i, ..., s_j\}$ contains a subset of sentences from the original document set D.

The system has to make a decision on which sentences to choose as part of summary from all k sentences. As mentioned above, the action space consists of all the possible actions that the system can take. In sentence extraction scenario system can pick any of the k sentences present in the documents set towards summary. Hence the action space is the entire sentence collection and the system has to make a decision about which sentence to pick, so $\{s_1, s_2..., s_k\}$ are all the possible actions for Document set D. So as per the discussion in section 3.1 if we pick a particular sentences s_i which is a possible action here, there will be a loss function $L(s_i, D)$ associated with this action which will specify our decision preferences in terms of selection of sentences. In risk estimation formalism, the risk of picking a sentence s_i when the information available to system is D, is given as

$$Risk(s_i) = \int L(s_i, D)\pi(D)dD \tag{3.5}$$

The expected risk of picking a sentence is measured in terms of a loss function $L(s_i, D)$, The loss function $L(s_i, D)$ measures the consequences of acting as if the true value of the quantity of interest were sentence s_i when the actual parameter value is document set D. In other terms its measure of regret of using this sentence instead of whole document set. Prior probability of document set $\pi(D)$ will be constant across the sentences and can be ignored. So the risk of picking a sentence is directly proportional to the amount of information lost when we pick a particular sentence s_i to represent the whole document cluster D. We can infer the estimates of s_i and D based on the data available i.e. the term space. The optimal sentence s^* will be the one with minimum expected risk.

$$s^* = arg_k min Risk(s_i) \tag{3.6}$$

So sentences with minimum risk is optimal to be chosen as part of summary. We consider sentence selection is a sequential decision making process, and pick sentences towards summary in the order of their risk value.

As discussed in section 3.1.1, in an inferential context we are interested in the discrepancy between the models labelled by the true value of the parameter and its estimate for loss function measurement. i.e. intrinsic loss function should be used. The loss function $L(s_i, D)$ will be used to measure the discrepancy between the models of s_i and D i.e. between M_{s_i} and M_D. respectivly. As discussed in section 3.2 we can use unigram langugae model to estimate $P(w|M_{s_i})$ and $P(w|M_D)$.

So an Intrinsic loss function of the form

$$L(s_i, D) = L\{P(w|M_{s_i}), P(w|M_D)\} \tag{3.7}$$

will be used to estimate the risk of picking a sentence.

3.4 Sentence Selection based on Information Loss

To measure the risk of infromation loss in picking a sentence we need a sentence representation, document representation, and a loss function to predict document from sentence model. The representation problem is thus equivalent to that of model estimation and a loss function is the invariant loss function.

We model the sentence selection process into four basic components: (1) View sentence as an observation from a probabilistic sentence model. (2) View document as an observation from a probabilistic document model. (3) Use of Intrinsic Loss function between sentence and document model to measure the loss associated with a sentence to represent the document cluster. (4) Consider sentences with minimum information loss as candidate for summary generation process.

3.4.1 Document and Sentence Representation

As discussed above we need a probabilistic represenation of document and sentence. We use simple unigram langauge model for this purpose as explained in section 3.3. For the document cluster D, we estimate $P(w|M_D)$,

$$P(w|M_D) = \frac{tf(w, D)}{|D|} \tag{3.8}$$

where $tf(w, D)$ is the frequency of word w in the document D and $|D| = \sum_w D(w)$ is total number of times all words occur in the document set D , it is essentially the length of the document cluster D.

Sentences also modeled in the same manner.

$$P(w|M_S) = \frac{tf(w, S)}{|S|} \tag{3.9}$$

here $tf(w, S)$ is the frequency of word w in the sentence S and $|s| = \sum_w d(w)$ is the sentence's length.

Smoothing of Probability Estimation: Similar to language models for speech recognition, these estimates can be smoothed for improving the probability estimates of a document model using various smoothing mechanism. Smoothing functions have been studied extensively for all the language modeling task, and it's been very effective in information retrieval research. Smoothing of probability estimates has a major impact on IR system performances. Here we will explore the effect of two very popular and effective smoothing functions (Jelinek-Mercer, Dirichlet prior) [40] to smooth document probabilities and later in the evaluation section will analyze the effect of this on summarization performance.

A simple yet effective smoothing procedure, which has been successfully applied in IR is Jelinek-Mercer smoothing. Jelinek-Mercer method involves a linear interpolation of maximum likelihood model with the collection model, using a coefficient λ to control the influence of each model.

$$P'(w|M_D) = (1 - \lambda)P(w|M_D) + \lambda P(w|C) \tag{3.10}$$

The probability of sampling a term w from document model $P(w|M_D)$ is estimated from the document cluster using a maximum likelihood estimator as described in equation 3.8. This estimate

is interpolated with the marginal $P(w|C)$ which is computed on a large background corpus.

$$P(w|C) = \frac{tf(w, C)}{|C|}$$

Dirichlet prior is an effective smoothing technique for text-based applications, in particular information retrieval.

$$P^{''}(w|M_D) = \frac{|D|}{\mu + |D|} P(w|M_D) + \frac{\mu}{|D| + \mu} P(w|C) \tag{3.11}$$

where μ is a smoothing parameter and $P(w|C)$ is same background model metioned above. Choice of an appropriate value for μ is a tuning stage in the use of language models. In principle the background model could be any source of typical statistics for components. Intuitively it makes sense to derive the model from other documents of similar type; in attributing newswire articles. As background model, we use the aggregate of all known documents, including training and test, as this gives the largest available sample of material.

3.4.2 Sentence Expected Risk

Risk of picking a sentence is proportional to loss of information when we consider sentence as a unit to represent the whole document cluster. We measure this according to the information theoretic intrinsic loss function relative entropy. It measures how much information is required to reconstruct the whole document cluster D from sentence distribution s, which is same as the loss of information when we decompose a document set D to sentence s. We measure relative entropy(RE) of the sentence model with the document model.

$$Risk(s_i) \propto L(s_i, D) \propto L\{P(w|M_{s_i}), P(w|M_D)\} \propto RE\{P(w|M_{s_i}), P(w|M_D)\}$$

$$RE(P(w|M_{s_i}), P(w|M_D)) = \sum_w p(w|M_{s_i}) log \frac{p(w|M_{s_i})}{p(w|M_D)} \tag{3.12}$$

The computation of the above formula involves a sum over all the words in D that have a non-zero probability according to $P(w|M_{s_i})$, because if there is a term $w \in D$ not in s_i, its contribution will be zero in the sentence information loss computation.

Each sentence s_i gets a expected risk $Risk(s_i)$ according to its relative entropy in comparison to document cluster D. The contribution of any term w in sentence information loss can be defined as how much we lose information by assuming the term is drawn from the sentence model instead of the document model. The Risk of sentence is total regret over all the words present in sentence.

In information theory, - log $P(w|M_D)$ is the amount of information contained in the word w under the document distribution, or equivalently, the minimum number of bits it takes to encode the word w in the optimal code based on the the distribution D. The relative entropy function measures the difference between the average number of bits to encode an word w when the true probability distribution is document distribution and the optimal code based on the distribution s is used. It is a measure of the regret we have at using the distribution s to define our text, instead of the optimal distribution D. In other words it is the regret of using a sentence to convey the information instead of whole document. The loglikelihood ratio plays a fundamental role to estimate loss for the particular sentence.

From the properties of relative entropy, the risk $Risk(s_i)$ is a nonnegative number for every i, and is 0 only when the estimated sentence distribution is the same as the document distribution. The goal of sentence extraction process is to pick sentences with the risk $Risk(s_i)$ as small as possible from all i.

As discussed above, relative entropy function estimates the amount of information required to reconstruct the original source document set. We try to approximate to the large document set D with a small sentence s_i, the idea is to view summarization as method to approximate a large document set by a smaller summary with minimum information loss. We try to approximate a document set by a small sentence set, or approximating large text with small text. On the whole the candidate of summary act as a surrogate for document set in a larger inference process. A large document set is divided into smaller text units (sentences), each of which is approximated variationally, yielding an overall variational approximation to the whole document set. This decomposition strategy leads us directly to a new sentence extraction algorithm. Furthermore, by considering a wider variety of divergence measures, we analyze different complexity and performance goals in next section.

3.4.3 Different Divergence Measures and Loss Function

In the sentence extraction framework we proposed to use an intrinsic loss function to measure the discrepancy between sentence and documents probabilistic models. We use relative entropy loss function as an intrinsic loss measure which is also known as directed logarithmic divergence for the purpose of comparing sentence and documents models and measure the information loss associated with sentence. Relative entropy comes from the family of information theoretic divergence measure and there are several divergence measures that have already been proposed. This opens up the possibility to include other divergence measures in sentence extraction framework and compare their effect on the performance. A divergence measure satisfies certain intuitive conditions which are satisfied by an entropy measure. The basic condition is that a measure of divergence should be nonnegative real valued function defined on the space of probability distributions which reflects the differences between the individuals within a population. The condition is a natural one since a measure of diversity should be nonnegative and should be value zero when all the individual probabilities are identical.

Information Theoretic Measures

Here we discuss some of the popular entropy based divergence measures that can be used to estimate loss between sentence and document model, and later in the experiment section of chapter 4 we analyze the performance of these different measures.

- Jensen Differencen Divergence Measure: Jensen Differencen Divergence Measure was given by Burbea and Rao [52, 53]. It is also known an Information radius [51]. When P and Q are two probability distributions, it is given by

$$R(P,Q) = H\left(\frac{P+Q}{2}\right) - \frac{H(P)+H(Q)}{2} \tag{3.13}$$

  This is also related to relative entropy:

$$R(P,Q) = \frac{1}{2}\left[RE\left(P, \frac{P+Q}{2})\right) + RE\left(Q, \frac{P+Q}{2})\right)\right] \tag{3.14}$$

Jensen Differencen Divergence Measure For the purpose of sentence risk estimation in summarization:

$$R(S,D) = \sum_{w} \left[\frac{p(w|M_{S_i}) log p(w|M_{S_i}) + p(w|M_D) log p(w|M_D)}{2} \right]$$
$$+ \sum_{w} \left[\left(\frac{p(w|M_{S_i}) + p(w|M_D)}{2} \right) log \left(\frac{p(w|M_{S_i}) + p(w|M_D)}{2} \right) \right] \quad (3.15)$$

- Jeffreys divergence: As we discussed earlier in section 3.2 that relative entropy is not a symmetric measure. Jeffrey [50] proposed as a symmetric version of relative entropy metric $J(P,Q) = RE(P,Q) + RE(Q,P)$ as a measure of divergence between two probability distributions. It is known as J-divergence in litrature.

$$J(S,D) = \sum_{w} \left(p(w|M_{S_i}) log \frac{p(w|M_{S_i})}{p(w|M_D)} \right) + \left(p(w|M_D) log \frac{p(w|M_D)}{p(w|M_{S_i})} \right) \quad (3.16)$$

Burbea and Rao [53] and Sgarro [55] established an enquality between J and R divergence. $J(P,Q) \geq 4R(P,Q)$

- Lin-Wong divergence: Lin and Wong [56] introduced following divergence measure $LW(P,Q) = RE(P, \frac{1}{2}P + \frac{1}{2}Q)$

$$LW(S,D) = \sum_{w} p(w|M_{S_i}) log \left[\frac{p(w|M_{S_i})}{\frac{1}{2}p(w|M_{S_i} + \frac{1}{2}p(w|M_D)} \right] \quad (3.17)$$

Lin and wong have shown various enqulalities. $LW(P,Q) \leq \frac{1}{2}RE(P,Q)$; $LW(P,Q) \leq 1$;

Other Populer (non-entropical) Divergence measures

- Hellinger loss: Hellinger loss [48] is an intrinsic loss function but not from the family of entropical loss (no an infomation theoretic measure). It is given by

$$HL(S,D) = \frac{1}{2} \sum_{w} \left(\sqrt{\frac{p(w|M_{S_i})}{p(w|M_D))}} - 1 \right)^2 \quad (3.18)$$

- Triangular Discrimination [49].

$$T(S,D) = \sum_{w} \frac{p(w|M_{S_i} - p(w|M_D)^2}{p(w|M_{S_i} + p(w|M_D)} \tag{3.19}$$

- Arithmetic-Geometric Divergence [55].

$$AG(S,D) = \sum_{w} \left(\frac{p(w|M_{S_i} + p(w|M_D)}{2} \right) log \left(\frac{p(w|M_{S_i} + p(w|M_D)}{2\sqrt{p(w|M_{S_i})p(w|M_D)}} \right) \tag{3.20}$$

- Symmetric Chi-square Divergence [49].

$$\psi(S,D) = \sum_{w} \frac{p(w|M_{S_i} - p(w|M_D)^2 p(w|M_{S_i} + p(w|M_D)}{p(w|M_{S_i})p(w|M_D)} \tag{3.21}$$

χ^2-divergence [54]

$$\chi^2(S,D) = \sum_{w} \frac{p(w|M_{S_i} - p(w|M_D)^2}{p(w|M_D)} \tag{3.22}$$

3.5 Summary Generation

For summary generation, we select sentences with minimum risk to produce summary while keeping the redundancy minimum. Having identified the sentences which are the most significant based on their risk value, the redundancy removal step attempts to reduce the size of this set without any reduction in information content. This is achieved by removing the sentences which duplicate information given in other sentences. The methods we used to perform simple redundancy removal are given in next section. With an optimal set of sentences produced after redundancy removal, the extracted sentences are reconstituted into a coherent summary in the text reformulation stage to establish a chronology.

3.5.1 Redundancy Identification

The sentence selection method adds sentences with minimum risk, one by one, within the appointed length of summary. In document set, the important information may be repeated more

than once, and a lot of the same or similar sentences may appear in the documents. So sentence selection should not only consider the score of the sentence but also measure its redundancy with the already selected sentences. After ranking sentences according to their risk, it first selects the sentence with minimum risk to the summary. As long as the summary is under the appointed summary length, the next sentence with least risk value is added to the summary if its redundancy with the already selected sentences in the summary is lower than the threshold. For redundancy identifications, we use the measure of number of terms overlapping between the already generated summary and the new sentence being considered.

The redundancy is measured as follows:

$$Redundancy(S_i, S) = \frac{|Term(S_i) \bigcap Term(S)|}{|Term(S_i)|} \tag{3.23}$$

where Term(Si) and Term(S) are the set of terms in Si and set S after stopword removel and stemming. Sentences are stopped and stemmed using the Porter algorithm and a count of all the common words in the sentences is calculated.

If the value of Redundancy is high, then it indicates that the sentence contains too many common terms with the already selected sentences, and it is irrelevant to add this sentence to the summary. We observed that a 40 percent term overlap between sentences is a good heuristic to estimate redundancy and hence used this redundancy threshold. So if there is a 40 percent overlap between the sentences already selected as part of summary and a new sentence, then that new sentence will not be a part of summary. We will give detail about the threshold selection in experiment section.

3.5.2 Sentence Organization

Sentence ordering is required to impose a coherent structure on a summary extracted from multiple sources. While it is trivial to order sentences from a single document (just use original ordering), this is not the case with multiple sources. The summaries need to reflect the changes to the story with respect to time. Hence, once sufficient number of sentences are picked to make the required length of summary, we arranged them based on chronological ordering (between documents i.e.

Algorithm 1 Summary Generation Steps

Step 1: Identify sentence boundaries in the given set of documents to decompose the document set D into individual sentences and form the candidate sentence set S=$\{s_i|i = 1, 2, .., n\}$.

Step 2: For each sentence $s_i \in S$ compute its expected risk value $Risk(s_i)$ using proposed mechanism, then sort the sentences in ascending order based on their risk.

Step 3: Select sentence s_i with minimum $Risk(s_i)$, and move it to the summary set F and remove it from S.

Step 4:

while $|F|$ <required summary length **do**

Pick the next sentence s_k with minimum $Risk(s_k)$ from set S

if term overlap between F and $s_k < r$ where r is redundancy threshold **then**

add s_k to F, remove s_k from S

else

remove s_k from S

end if

end while

Step 5: Arrange the sentences in F in chronological order i.e. in the order found in the source documents or based on the order of occurrence if they are from the same document.

based on the time stamp) and order of occurrence (within the document). Thus, sentences coming from different document will be ordered based on their source documents date of publication and if two sentences originate from the same document their original order in the source document will be considered to generate the final summary. Any additional words than the required length of summary are truncated. Algorithm 1 shows the operation flow of summary generation process. This sentence organization method is applicable to only news domain, we may need alternative models for sentence organization for a good coherence in the other domains.

3.6 Summary

In this chapter, we present the Information Loss framework for document summarization. Since the formulation is based on decision theory and probabilistic language modeling, we first gave a brief introduction to both. We discussed statistical decision theory which provides a theoretical foundation to deal with problems of action and inference under uncertainty. Then we discussed how loss functions have been used in making a decision, which represents the loss associated with an action being ”wrong” as a function of a measure of the degree of wrongness. Then we discuss why we use intrinsic loss function instead of conventional loss functions in our setting and conclude relative entropy as a suitable intrinsic loss function. After this we formalized the sentence extraction summarization process as a decision making task, where each sentence is a possible action that can be taken, and system has to make a decision on which sentences to pick as part of summary. Through this formulation we defined the extraction process into three basic components, which is a sentence representation, document representation, and a loss function to predict document from sentence model. We used language modeling technique for probabilistic representation of document and sentence, and an information theoretic intrinsic loss function (relative entropy) which measures the loss in terms of modeling how bad a sentence distribution is in predicting document distribution. After discussing our approach of sentence loss estimation, we discuss how to tackle the problem of redundancy for summary generation. Our redundancy mechanism is a term

overlap phenomena. If a sentence contains too many common terms with the already selected sentences, it is irrelevant to add this sentence to the summary. And after this content selection process we reorganize the sentences based on chronological ordering (between documents) and order of occurrence (within the document) to improve readability.

Chapter 4

Experiments and Evaluation

4.1 Evaluation

In order to evaluate the performance of our approach, we use three data sets that have been widely used in recent multi-document summarization evaluations: DUC-2004, MSE-2005, DUC 2007. For evaluation we used the automatic summary evaluation metric, ROUGE [7] which is the standard way of evaluation of summaries. ROUGE is a recall based metric for fixed-length summaries which is based on n-gram cooccurrence. It measures summary quality by counting overlapping units such as the n-gram, word sequences and word pairs between the candidate summary and the reference summary(human written summaries) . We show three of the ROUGE metrics in our experimental results: ROUGE-1 (unigram-based), ROUGE-2 (bigram-based), and ROUGE-SU4 (skip bigram) metric. [1]

4.2 Evaluation Metric

ROUGE [7] is an intrinsic evaluation metric to automatically score the peer summaries based on pairwise comparison with the reference summaries. ROUGE evaluates each peer summary

[1]ROUGE version 1.5.5, with arguments -n 2 -x -m -2 4 -u -c 95 -r 1000 -f A -p 0.5 -t 0 -d -e

based on its n-gram similarity with the reference summaries. Different ROUGE measures were proposed [49] and shown to be correlating well with manual evaluations. ROUGE-N is an ngram recall between a candidate summary and a set of reference summaries. ROUGE-N is computed as follows:

$$ROUGE - N = \frac{\sum_{S \in ReferenceSummaries} \sum_{gram_n \in S} Count_{match}(gram_n)}{\sum_{S \in ReferenceSummaries} \sum_{gram_n \in S} Count(gram_n)} \tag{4.1}$$

Where n stands for the length of n-gram, $gram_n$, and $Count_{match}(gram_n)$ is the maximum number of n-grams co-occurring in a candidate summary and a set of reference summaries. From the expression it is clear that it is a recall oriented measure. When multiple reference summaries were used, a pairwise summary-level ROUGE-N between a candidate summary s and every reference, r_i, in the reference set is computed. Then the maximum of pairwise summary-level ROUGE-N scores is treated as the final multiple reference ROUGE-N score. ROUGE-1 and ROUGE-2 are being used by the current summarization research community. ROUGE-SU is another metric which is based on the skip-bigram co-occurrence statistics. Skip-bigram is any pair of words in their sentence order, allowing for arbitrary gaps. Skip-bigram co-occurrence statistics measure the overlap of skip-bigrams between a candidate summary and a set of reference summaries. For example, the skip-bigrams identified for sentence “police killed the gunman” are “police killed”,“police the”,“police gunman”,“killed the”,“killed gunman” and “the gunman”. Given a reference summary X of length m and peer summary Y of length n and if SKIP2(X, Y) is the number of skip-bigrams between X and Y , then the F-measure is computed as follows:

$$R_{skip2} = \frac{SKIP2(X,Y)}{C(m,2)}$$

$$P_{skip2} = \frac{SKIP2(X,Y)}{C(n,2)}$$

$$ROUGE - S = F_{skip2} = \frac{(1+\beta^2) R_{skip2} P_{skip2}}{R_{skip2} + \beta^2 R_{skip2}}$$

Where β controls the relative importance of P_{skip2} and R_{skip2} and C is the combination based function. One problem for ROUGE-S is that it does not give any credit to a candidate sentence if

the sentence does not have any word pair co-occurring with its references. An extended measure called ROUGE-SU is created to take this into consideration.

4.3 Dataset

We used three different Datasets DUC2004, MSE2005 and DUC2007 to run and evaluate the performance of our system. DUC2004 is a widely accepted benchmark dataset to evaluate Generic Multi document Summarization performance, with most of the recently proposed systems using this dataset to evaluate the summaries. MSE2005 dataset has been used in Machine Translation and Summarization Workshop at ACL 2005. DUC 2007, is the most recent dataset in the series of multiple document summarizations evaluations of document understanding conference, is a Query-focused multi document summarization task. Although the primary task here is not of Generic Multi-document summarization, since each cluster may have documents relating to the same topic, a generic summarization is also expected to produce satisfactory results without using queries. In the following sections we describe each of these datasets and the provided baselines for evaluation, and then we present the comparison of our proposed system (RE Summarizer) with the top performing systems on these datasets.

4.3.1 DUC 2004

The Document Understanding Conference (DUC) has been carrying out large-scale evaluations of summarization systems on a common dataset since 2001. On average, over 25 different sites (International research groups) participate in this NIST-run evaluation each year and a lot of effort has been invested by the conference organizers to improve evaluation methods. DUC content evaluations are still based on a system summary comparision to human model summaries. However, in order to mitigate the bias coming from using gold-standards from only one person, different annotators create the models for different subset of the test data.

The generic multi-document summarization task in DUC 2004 (task 2) involves summarization

of 50 TDT (Topic Detection and Tracking) English clusters where each cluster has 10 news articles discussing the same topic.The task was to generate a 100 word summary for each cluster. Four different human judges produced model summaries (reference summaries) for any given cluster for comparison.

System	ROUGE-1	95% conf. interval
RE Summarizer	0.38664	0.37967 - 0.39372
peer65	0.37950	0.37345 - 0.38535
peer104	0.37115	0.36536 - 0.37691
peer35	0.37567	0.36908 - 0.38181
Coverage baseline	0.34542	0.33297 - 0.35769
Lead baseline	0.32100	0.31468 - 0.32777

Table 4.1: ROUGE-1 systems comparison on DUC 2004.

The proposed approaches are compared with the top 3 performing systems and two baseline systems (i.e. the lead baseline and the coverage baseline) on task 2 of DUC 2004. The top three systems are the systems with the highest ROUGE scores, chosen from all participating systems in the tasks of DUC 2004. The lead baseline and coverage baseline are two baselines employed in the multi-document summarization tasks at DUC. Lead baseline simply takes the first 100 words of the most recent news article in the document cluster as the summary. And the coverage baseline takes the first sentence from the first document, the first sentence from the second document, and the first sentence from the third document, and so on, until the summary reaches the length limit.

The first column of table 1 represents different summarization system, here RE Summarizer is our proposed summarization system which uses relative entropy based loss function for sentence extraction. The next three systems are top performing system on given dataset and last two systems are converge and lead baseline as explained above. Second column represents the ROUGE Perfor-

System	ROUGE-2	95% conf. interval
RE Summarizer	0.09423	0.08840 - 0.09986
peer65	0.09171	0.08733 - 0.09642
peer104	0.08489	0.08082 - 0.08903
peer35	0.08389	0.07888 - 0.08879
Coverage baseline	0.07452	0.06734 - 0.08208
Lead baseline	0.06375	0.05915 - 0.06852

Table 4.2: ROUGE-2 systems comparison on DUC 2004

System	ROUGE-SU4	95% conf. interval
RE Summarizer	0.13550	0.13063 - 0.14035
peer65	0.13238	0.12845 - 0.13628
peer104	0.12805	0.12447 - 0.13152
peer35	0.12907	0.12516 - 0.13273
Coverage baseline	0.11396	0.10781 - 0.12011
Lead baseline	0.10225	0.09877 - 0.10592

Table 4.3: ROUGE-SU4 systems comparison on DUC 2004.

mance of this system and the last column shows range of the scores with a 95 percent confidence.

We can see from The table 4.1,4.2 and 4.3 that our system outperforms the top performing systems and baseline systems on DUC 2004 tasks over all three ROUGE metrics. Our system is substantially simpler to other system still we are able achieve better performance. We compare the theoretical complexity of our system with these top performing systems and analyze our good performance in section 4.7 and 4.8. In table 4.4 we show an example of our system generated summary for topic d30001 which contains 10 news articles. We also provide two out of the for human written model summaries provided for evaluation.

4.3.2 MSE 2005

In 2005, a multi-document summarization task was conducted as part of the Machine Translation and Summarization Workshop at ACL. A total of 25 document sets containing a mixture of English and machine translated Arabic news to English were provided and 100 word summary needs to be generated. The news articles are generally shorter than those used in DUC-2004. Ignoring the potential mistakes introduced by the machine translator, we ran our systems without any modifications for this unusual setting.

As shown in Table 4.5, on this data set, our system performs better than the top performing systems on ROUGE-1 metric. Performance with other two metrics are comparable to top system. In table 4.6 we show an example of our system generated summary for topic 33001 which contains 10 news articles. We also give two out of the four human written model summaries provided for evaluation.

4.3.3 DUC 2007

Unlike DUC2004 and MSE 2005, DUC 2007 is not a Generic Multi-document summarization task. Instead it is a Query-focused multi document summarization task, with 50 English document clusters generated from AQUANT corpus, where each cluster has 25 news articles discussing the

Table 4.4: Example: System generated and Human written summaries from DUC 2004 dataset

Our system summary: Hun Sen's Cambodian People's Party And Ranariddh's FUNCINPEC Party Agreed To Form A Coalition That Would Leave Hun Sen As Sole Prime Minister And Make The Prince President Of The National Assembly. In A Long-elusive Compromise, Opposition Leader Prince Norodom Ranariddh Will Become President Of The National Assembly Resulting From Disputed Elections In July, Even Though Hun Sen's Party Holds A Majority Of 64 Seats In The 122-member Chamber. Opposition leaders Prince Norodom Ranariddh and Sam Rainsy, citing Hun Sen's threats to arrest opposition figures after two alleged attempts on his life, said they could not negotiate freely

Model summary 1: Prospects were dim for resolution of the political crisis in Cambodia in October 1998. Prime Minister Hun Sen insisted that talks take place in Cambodia while opposition leaders Ranariddh and Sam Rainsy, fearing arrest at home, wanted them abroad. King Sihanouk declined to chair talks in either place. A U.S. House resolution criticized Hun Sen's regime while the opposition tried to cut off his access to loans. But in November the King announced a coalition government with Hun Sen heading the executive and Ranariddh leading the parliament. Left out, Sam Rainsy sought the King's assurance of Hun Sen's promise of safety and freedom for all politicians.

Model summary 2: Cambodia King Norodom Sihanouk praised formation of a coalition of the Countries top two political parties, leaving strongman Hun Sen as Prime Minister and opposition leader Prince Norodom Ranariddh president of the National Assembly. The announcement comes after months of bitter argument following the failure of any party to attain the required quota to form a government. Opposition leader Sam Rainey was seeking assurances that he and his party members would not be arrested if they return to Cambodia. Rainey had been accused by Hun Sen of being behind an assassination attempt against him during massive street demonstrations in September.

System	ROUGE-1	ROUGE-2	ROUGE-SU4
RE Summarizer	.4392	.1602	.1860
peer28	.4342	.1603	.1862
peer29	.4231	.1435 .	.1696
peer30	.4188	.1397	.1690

Table 4.5: Comparison on MSE 2005

same topic. The participants were asked to generate a 250 word summary for each cluster using the given query. As each cluster had documents relating to the same topic, a generic summarization is also expected to produce satisfactory results without using queries.

In DUC 2007 two baselines are given. First is the leadline same as DUC 2004 which simply takes the first 250 words of the most recent news article in the document cluster as the summary. The second baseline is the system CLASSY, a generic multi-document summarizer, the best performer in both DUC 2004(peer65) and MSE 2005(peer28). Both of these baselines are generic summaries. We compare our results with both these baselines on this dataset. Even with DUC 2007 our system gives better results than CLASSY as is the case with DUC2004 and MSE 2005. The results for DUC 2007 is shown in Table 4.7. In table 4.8 we show an example of our system generated summary for topic D0703A which contains 25 news articles. We also give two out of the four human written model summaries provided for evaluation.

One important thing to notice here is that our system performance with respect to top system gets better as the dataset size becomes large. In MSE 2005 dataset each document cluster contains 8-10 news articles. In DUC 2004 each document cluster contains 10 documents but each document size is larger as compare to MSE 2005. In DUC 2007 document clusters are of bigger size, here each cluster contains 25 documents. The performance of our system with respect to the previous top system on MSE 2005 is comparable but performance gap is marginal. In DUC 2004 our system

Table 4.6: Example: System generated and Human written summaries from MSE 2005 dataset

Our system summary: Canadian Federal Police Announced The Arrest Of Two Sikh Today, Friday, Killing 239 People In The Explosion Of A Boeing 747 Indian Airlines Flight Off The Coast Of Ireland, 32 June 1985 Had Been Taken Off From Canada. Two Men Of Sikh Origin Were Arrested And Charged Friday With Killing 329 People In The 1985 Bombing Of An Air India Jumbo Jet Near Ireland And Two Baggage Handlers Killed By Another Bomb In Tokyo. Air India Flight 182 from Montreal to New Delhi, with a planned stop in London, went down off the coast of Ireland on June 23, 1985

Model summary 1: The Canadian Federal Police announced the arrest of two Vancouver-based Sikhs on October 27, 2000. They were charged with the deaths of 331 people, 329 of whom were killed when Indian Airlines Flight 182 exploded off the coast of Ireland, and two of whom were baggage handlers killed in an airplane bombing in Tokyo. Both incidents occured on June 23, 1985. The individuals arrested were Ripudaman Singh Malik, 53, and Ajaib Singh Bagri, 51. Canadian authorities at the time of the bombing had suspected Sikh extremists of attempting to avenge the 1984 raid on their sacred Golden Temple in Amritsar.

Model summary 2: Two men of Sikh origin, Singh Malik, 53, and Ajaib Singh Bagri, 51, were arrested and charged Friday with killing 329 people in the June 23, 1985 bombing of an Air India jumbo jet near Ireland. They were also charged with killing two baggage handlers by another bomb in Tokyo the same day. This is Canada's biggest mass murder case. According to the Royal Canadian Mounted Police, this was the most complex investigation in the force's history. Canadian investigators have long suspected Sikh militants of planting the bombs in revenge for India's 1984 raid on the Golden Temple in Amritsar.

System	ROUGE-1	ROUGE-2	ROUGE-SU4
RE Summarizer	.4267	.1096	.1628
CLASSY	.4005	.0938	.1464
Lead baseline	.3347	.0603 .	.1050

Table 4.7: Comparison on DUC 2007

significantly outperforms next best system and in DUC 2007 the scores are even higher. This signifies that when the documents size becomes larger i.e., when there is more challenging task to extract sentences from large diverse collection, our system performs better.

4.4 Evaluating Different Divergence Measures as Loss Function

In this section we evaluate the performance of different loss functions that can be used in the proposed framework presented in section 3.5.3. All the loss functions perform better than the baseline systems. Table 4.9,4.10,4.11 shows that RE summarizer performs better than other measures, which shows its effectiveness as a loss function as compare to other measures. Relative entropy has the classical information theoretical background and relevance of log scoring(discussed in section 3.2). Jensen, Lin-Wong and Jeffreys divergence as a loss function perform well and scores are consistantly good and comparable to RE summarizer. Jensen, Lin-Wong and Jeffreys divergence measure as a loss function perform well and scores are consistently good and comparable to RE summarizer. All of these three measures are entropic measures and that could be the reason of their effectiveness while measuring the loss of information. All three of these of these measures along with relative entropy measure are special cases of Renvi relative information [3].

Jeffreys divergence is actually a symmetric version of relative entropy and achieves good per-

Table 4.8: Example: System generated and Human written summaries from DUC 2007 dataset

Our system summary: Italy And France Have Adopted The Euro, As The European Union's New Single Currency Is Known. Dkt-cb Thailand Is Considering Using The European Single Currency, The Euro, In The Country's Foreign Reserves, The Nation Reported Tuesday. European Central Bank President Wim Duisenberg Said Thursday That Europe's New Single Currency, The Euro, Won't Compete With The U.S. Dollar As The Currency Of Choice For Foreign Reserves. European Union's Single-currency Plan Will Take Effect In 11 Of Its Member States On January 1, 1999, When The Euro Will Become A Currency In The Financial Sense, But Will First Be Used Only In Cashless Tradings. (gb) The Mass Printing Of The Banknotes Of The Single European Currency, The Euro, Would Be Started At The Beginning Of 1999, The European Central Bank (ECB) Announced In Frankfurt Friday. Private Investors Are Set To Benefit From The 1999 Introduction Of The Single European Currency, The Euro, Thanks To Expanded Investment Opportunities, Reduced Currency Risks, And Rising Competitions, The German Central Bank Bundesbank Said Monday. Trading In Europe's New Single Currency Will Begin On Monday As A United States Bank And A Dutch One Have Offered To Quote Prices In The Euro. France Has Already Struck One Billion Coins Of The European Single Currency Euro, Which Will Be Launched By January 1999, Reported Agence France-Presse (AFP) On Thursday. EU finance ministers, who fixed the exchange rates between the euro and the national currencies of the 11 as the final step before creating the new currency at midnight,

Model summary 1: European Union (EU) nations agreed that a single currency (the Euro) will go into effect on January 1, 1999. Polls indicate support but widespread skepticism remains. Eighty percent in six countries say they are not well informed. Some economists worry about loss of financial sovereignty; others worry about rising unemployment and interest rates. Proponents say the Euro will guarantee currency stability, lower interest rates and contribute to the unity of the EU. Belgium, Germany, Spain, France, Ireland, Italy, Luxemboug, the Netherlands, Austria, Portugal and Finland are founding members of the Euro club. Britain and Demark have opted out while Greece and Sweden have been judged economically not ready to join. The Euro will rival the U.S. dollar as an international currency but will not replace the U.S. dollar as the choice for foreign reserves. Initial transactions will be cashless. Bank notes and coins will become legal tender January 1st 2002 while national currencies will stop circulating by July 1st 2002. France, Finland, Belgium and Spain have started production of Euro's. A total of 70 billion coins should be issued to replace the national currencies. The Vatican, San Marino and Monaco are entitled to use the Euro as their official currency but cannot issue any currency unless they agree to EU conditions. The design of the Euro is required to include five languages, the symbol of the EU (12 stars in a ring) and will feature bridges, windows and doorways in various European styles. "EUR" will be the currency code.

Model summary 2: The Euro was scheduled to be launched on January 1, 1999, and preparations for its introduction were well underway three years before that date. By April 1996 a consultative group was in place to design the new currency and a year later a new currency code was issued to facilitate technical preparations. The euro was to be used in cashless trading as of January 1, 1999, but the actual currency was not to go into use until January 1, 2002, with present currencies' ceasing to be legal tender six months later. By mid-November 1998 some countries had begun to produce coins. The announcement by two banks in May 1998 that they would quote prices in euro seven months before its official adoption demonstrated the acceptance of the euro by financial markets. In early 1996 many EU countries feared that the Euro would cause them to loose financial sovereignty; however, by late 1997 Germany was encouraging its companies and investors to welcome the euro. Reaction to its introduction was positive elsewhere as well. China welcomed the move; Bulgaria worked to tie its currency to the euro; and Thailand considered using it in its foreign reserves. By 21 November 1998 Indian banks were preparing for euro transactions and by mid-December Romania's banking system was ready for the new currency. On 31 December Bulgaria announced that its foreign currency reserves would be backed up by the euro instead of the German marks to which it had formerly been tied.

System	ROUGE-1	ROUGE-2	ROUGE-SU4
RE summarizer	.3866	.0942	.1355
Hellinger	.3622	.0834	.1252
AG divergence	.3546	.0784	.1204
T Discrimination	.3647	.0803	.1250
Jensen divergence	.3767	.0912	.1272
Jeffreys divergence	.3822	.0938	.1344
Lin-Wong divergence	.3802	.0924	.1337
Chi-square	.3655	.0821	.1224
Symm Chi-square	.3673	.0834	.1242

Table 4.9: Different Loss Functions performance on DUC 2004

System	ROUGE-1	ROUGE-2	ROUGE-SU4
RE summarizer	.4352	.1602	.1868
Hellinger	.4067	.1096	.1428
AG divergence	.3805	.0938	.1364
T Discrimination	.3347	.0603	.1050
Jensen divergence	.4252	.1503	.1702
Jeffreys divergence	.4372	.1600	.1808
Lin-Wong divergence	.4264	.1524	.1742
Chi-square	.4186	.1357	.1610
Symm Chi-square	.4257	.1406	.1628

Table 4.10: Different Loss Functions performance on MSE 2005

System	ROUGE-1	ROUGE-2	ROUGE-SU4
RE summarizer	.4267	.1096	.1628
Hellinger	.3907	.0873	.1290
AG divergence	.3647	.0763	.1010
T Discrimination	.3847	.0863	.1210
Jensen divergence	.3985	.0948	.1524
Jeffreys divergence	.4367	.1098	.1602
Lin-Wong divergence	.4085	.0988	.1564
Chi-square	.3747	.0803	.1110
Symm Chi-square	.3867	.0842	.1202

Table 4.11: Different Loss Functions performance on DUC 2007

formance which is very close to RE summarizer. With DUC 2007 dataset it actually performs slightly better (not significant) than RE summarizer on ROUGE-2 and ROUGE-SU4 metrics. Out of all these divergence measures relative entropy(RE summarizer) proved to be most simple, consistent and effective measure to estimate the information loss in document summarization scenario.

4.5 Redundancy Threshold

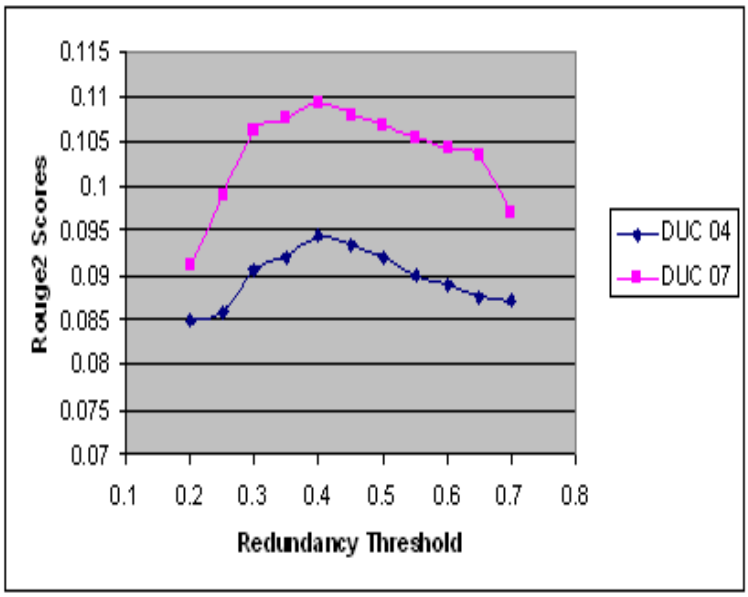

Figure 4.1: Summarization performance (ROUGE-2) with respect to different values of redundancy threshold

We experimented with different values of redundancy threshold on different datasets. Fig 4.1 shows the system performance on ROUGE2 metric with different values of redundancy threshold on DUC 2004 and DUC 2007 datasets. The system performance is optimum when threshold is set to .4 and quite constant between .4 and .5, same result reflects on ROUGE-SU4 metric. The system performance on DUC 2007 dataset is more sensitive to redundancy threshold as compared to

DUC 04. It is probably because document cluster size in DUC 07 is larger (25 documents related to same topic), so there is higher possibility of redundant information availability. The overall results conclude that a 40 percent term overlap between sentences is a good heuristic to estimate redundancy i.e. if there is a 40 percent overlap between the sentences already selected as part of summary and a new sentence, then that new sentence should not be a part of summary.

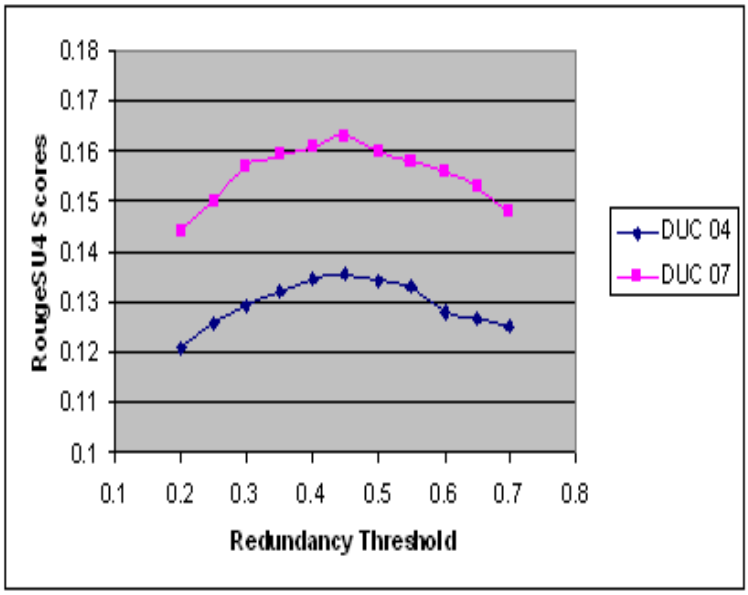

Figure 4.2: Summarization performance (ROUGE-SU4) with respect to different values of redundancy threshold

4.6 Effect Of Smoothing

We experimented with different values of smoothing parameter on different datasets. First we consider the performance of our proposed system(RE Summarizer) when we use Jelinek-Mercer smoothing in estimation of document model (equation 3.9). Fig 4.3 shows the system performance on ROUGE-2 metric with different values of smoothing parameter λ on DUC 2004 and DUC

2007 dataset. On DUC 2004 dataset as we increase the value of lambda the performance increases proportionally, but the improvement is very low and can not be considered significant. System achieves best performance for $\lambda = 0.5$, after which smoothing has negative impact on the performance and scores get down considerably when we increase $\lambda > 0.6$. On DUC 2007 dataset the effect of smoothing is even lesser as compared to DUC 2004, with an optimal value for smoothing parameter λ of 0.4. The maximum increment in performance for DUC 2007 is .0016 (.1096 to .1112) as compared to .0021 (.09423 to.0964) for DUC 2004 dataset. Fig 4.4 shows the system performance on ROUGE-SU4 metric, with similar patterns, system achieves highest performance for λ equal to 0.5 and 0.4 for DUC 2004 and 2007 respectively.

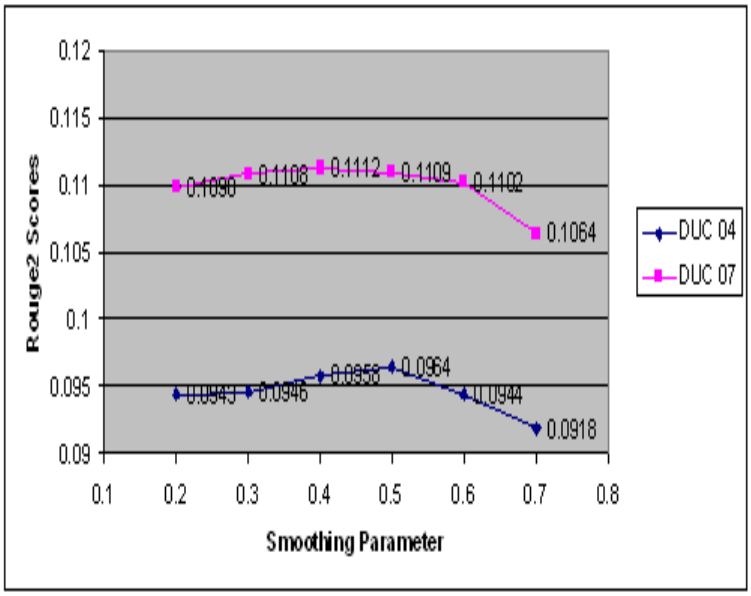

Figure 4.3: Summarization performance (ROUGE-2) with respect to different values of Jelinek-Mercer smoothing parameter

Dirichlet prior smoothing: Here we consider the performance of system when we use Dirichlet prior smoothing in estimation of document model (equation 3.10). Fig 4.5 shows the system performance on ROUGE-2 metric with different values of smoothing parameter μ on DUC 2004 and

DUC 2007 dataset. On DUC 2004 dataset, as we increase the value of lambda the performance increases proportionally, but the improvement is very low and can not be considered significant. System achieves best performance for $\mu = 100$, after which smoothing has negative impact on the performance and scores get down considerably when we increase $\mu > 1000$. On DUC 2007 dataset the effect of smoothing is even lesser as compared to DUC 2004, with an optimal value for smoothing parameter μ of 1000. The maximum increment to performance for DUC 2007 is .0019 (.1096 to .1115) as compared to .0026 (.09423 to.0968) for DUC 2004 dataset.

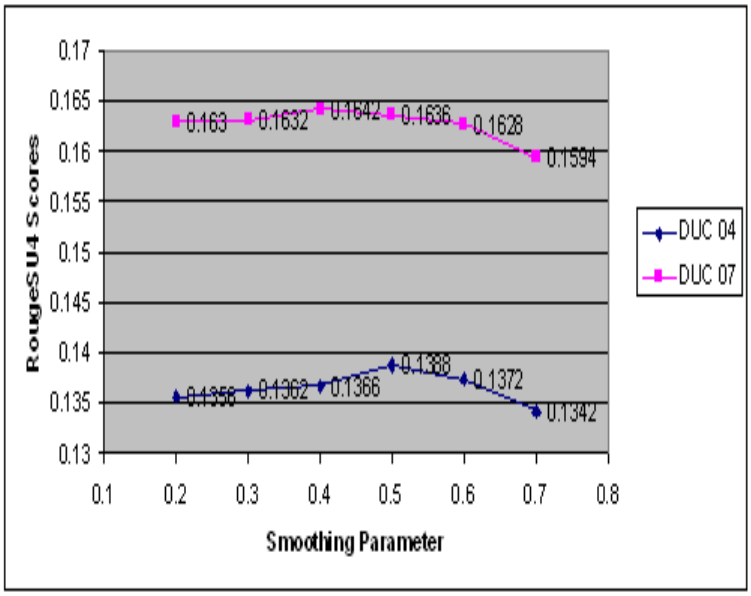

Figure 4.4: Summarization performance (ROUGE-SU4) with respect to different values of Jelinek-Mercer smoothing parameter

We can notice here that the optimal value of smoothing parameter is different for different dataset, and there is no common ideal value for this parameter suiting all the datasets. Hence for new data, it has to be purely empirical. The performance increase using smoothing functions is not very significant, and a wrong value can decrease the performance considerably. As compared to this, the smoothing functions have great impact on the performance of information retrieval

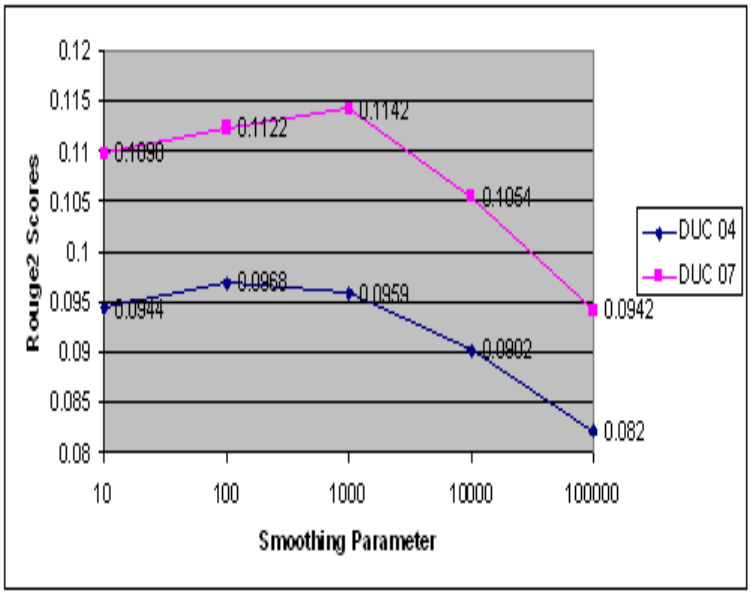

Figure 4.5: Summarization performance (ROUGE-2) with respect to different values of Dirichlet prior smoothing parameter

systems [24, 25]. This reduced impact of smoothing function shows that in summarization, the contextual information has more importance i.e. term probabilities with in the document cluster have enough information to define the context.

For datasets DUC 2004 and MSE 2005 smoothing functions has more impact on the performance as compared to DUC 2007. This may be due to the size of dataset. DUC 2004 and MSE 2005 have smaller document sets; thus the outside information influences the term probabilities more. In DUC 2007 each document set contains 25 documents, so it has more contextual information present in the document itself. Thus the smoothing of extra source doesn't impact the performance much. The experiment concludes that the smoothing function provides only a meager improvement on the performance of the summarization system. These functions can therefore be ignored for real time system considering the amount of complexity they add up to the system. Specifically, these functions must be trained to estimate parameters with some background corpus.

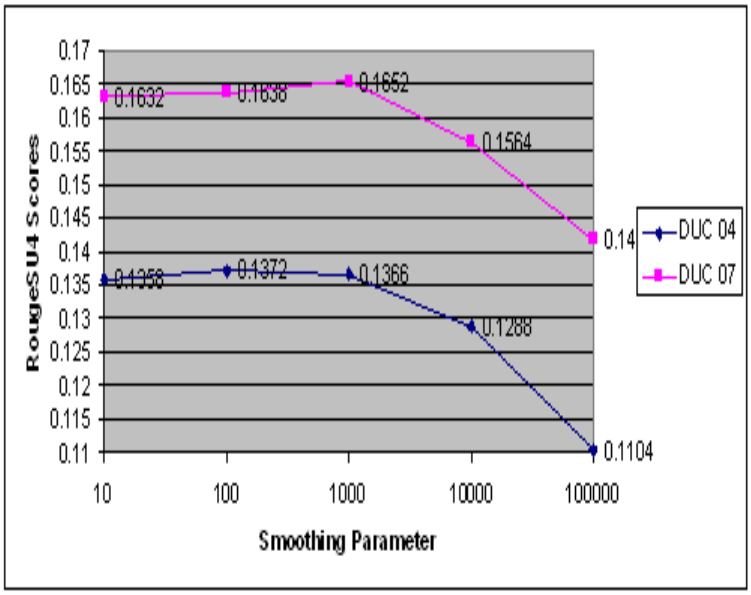

Figure 4.6: Summarization performance (ROUGE-SU4) with respect to different values of Dirichlet prior smoothing parameter

4.7 Related work comparision

CLASSY [26, 27] the previous best system on DUC 2004 and MSE 2005 consists of two core components - a Hidden Markov Model (HMM) for selecting sentences from each document and a pivoted QR algorithm for generating a multi-document summary. The HMM has two kinds of states, which correspond to summary and non-summary sentences in a single document on the basis of some signature terms. The model uses number of signature terms in each sentence as a feature. These terms are decided by the statistic computation similar to [29], derived based on a large set of documents in advance. In addition, the best number of the HMM states needs to be determined based on empirical testing, and the HMM model needs to be learned using training data. After applying the HMM, the top scoring sentences of each document form a weighted token-sentence matrix. A pivoted QR algorithm is then used for scoring and selecting sentences to form the output summary. In addition to these two core components, CLASSY also incorporates a linguistic component as a preprocessing stage to provide the summarization engine simplified

sentences as input.

Our Relative entropy based summarizer performs better than this the best performing CLASSY system. The results are very positive because the top performing system is supervised, and needs manually annotated training data, as well as background corpus. It also uses empirically set parameters (number of states for the HMM and topic signature word likelihood ration cut off). Finally, the theoretical complexities of top system make it difficult to understand of what makes the system work.

This HMM model based summarization, uses some signature tokens as one of the feature. These tokens are the terms that are more likely to occur in the document than in the corpus. As opposed to this, in our approach, while comparing sentence and document cluster language models, we actually compare sentence relationship with whole document cluster rather than few topical signatures. Also it was empirically observed by us that in summarization task smoothing has very less impact on the performance of the system. So the effectiveness of our approach basically depends on the comparison of sentence model against documents model and not on the comparison of corpus against document cluster. Above all our system is substantially simpler and light weight compared to CLASSY and is able to achieve better results.

As discussed in related work in 2.2.1 Radev's [13] proposed Mead summarization system uses several features for sentence extraction. Comparison of our results with Meads as reported on DUC 2004 shows that our system outperforms it by a significant margin using a single feature, i.e. term probability. Theoretically the centroid based feature of Mead can be compared with our approach. In centroid based method, the highly frequent terms in the document cluster are computed first to create a centroid feature vector containing a few highly frequent terms. Sentences are then scored based on its similarity with centroid feature vector using cosine similarity. As mentioned above, in our approach, while comparing sentence and document cluster language models, we actually compare sentence relationship with whole document cluster rather than few topical signatures or centroid themes. This is because we believe that a sentence is a function of whole document and its relationship with the entire document cluster should be considered while scoring. We follow a

”full coverage approach” to extract sentences that belong to the full concept-space of a document cluster. One fundamental difference with Radev’s work is that we use probabilistic model for representation of text as compared to Radev’s vector space theory.

One of the most recent systems proposed by Yih etc al. [32] also uses few signature terms to calculate sentence importance. As we discussed above our approach is different to signature based estimation, also it uses word positions information and a generative/decimating learning algorithm (machine learning) to combine frequency and position information. It also uses a search algorithm called stack decoder to search best sentence using frequency and position information. We compared complexity of our system with Classy system earlier, similar to that our system is light weight compared to [32] as we don’t use any position information or training data for sentence selection process, still we are able to achieve better results.

Our approach can be compared with all statistical features of earlier proposed summarization. In all such statistical approaches term frequency as one of the features along with several others determines sentence importance. However they vary in the way they use it for scoring. The effectiveness of a approach depends on how one uses the term feature to come up with a good composition function for scoring (For example in IR research term-frequency is the main parameter used for the last few decades). So the difference lies on how to exploit the information hidden in the term distributions. In proposed approach we consider sentence and documents as different text units and represent them with probabilistic distribution of terms. We then measure how bad a sentence probability distribution is in modeling the documents distribution using relative entropy. By doing this we are able to capture the amount of information loss when we pick a sentence as a single unit to represent the whole document cluster. No other previous approach is able to capture this.

4.8 Discussion

We proposed and evaluated a decision theoretic statistical summarizer. The summarizer we described is unsupervised and data driven, the measure it uses is information loss that takes only term probabilities as Input. The proposed extraction framework has simple comparative modeling mechanism, which allows us to examine the choice of a loss function for assigning a risk score to sentences. Different loss functions lead to differently performing summarizers, ranging from close to baseline to state of the art performance. Our results show that choosing relative entropy as a loss function gives the most balanced summarizer, which has very light weight computation and very good performance on different datasets. We also saw that redundancy identification is an important step for good summarizer as it improves the performance of summarizer effectively. Relative entropy based summarizer significantly outperforms most of other systems, and compares favorably with the best performing systems. In summary we have proposed a conceptually simple algorithm for generic multi document summarization that achieves the state of the art performance.

The effective performance of our approach shows that information loss depicts the quality of the summaries generated. Automatic Summarization is defined as a process whose goal is to produce a condensed representation of the content of its input for human consumption [9] . In automatic document summarization, input is documents which are viewed as information sources whose content reflects things in the world. So we can view summarization as a condensation or compression process of an information source. We know the compression or condensation process of any information source suffers from loss of information. In proposed approach while modeling the risk of picking a sentence towards summary, we are minimizing the loss of information. So we are able to generate summary which preserves the semantic informativeness of original source document i.e. the summary is informative and correlates well with the human written summaries.

With relative entropy metric we pick the sentences towards summary based on its capability to reconstruct the source document. In information theoretic terms, a summary could be viewed as

informative if it allows one to reconstruct the source document based on it. So if a human were asked to guess the content of the source text based on reading the summary, the best summary would be one which allowed the human to correctly guess the full text of the source document.

As explained above we approach the summarization process as a compression or condensation process of an information source. This may create confusion between summarization and text compression, as the objective looks quite similar. As per the definition given by bell 1990, Text compression aims at considering a text input, by treating input as code. The condensation process here involves taking advantage of the redundancy in the input. The compressed representation is intended for efficient storage and transmission among machines, rather than for human consumption. While in summarization process the condensed representation of the content is for human consumption. So we can say that compression is machine level condensation and summarization is human level condensation task. In compression we treat text as code but in summarization it is a semantic unit. Compression algorithms use basic binary level representation as compared to summarization algorithms which use term level representation.

4.9 Summary

We evaluated the performance of our system on three widely accepted multi-document summarization task Datasets viz., DUC 2004, MSE 2005 and DUC 2007. We used ROUGE as an intrinsic evaluation metric, to automatically score the peer summaries based on pair-wise comparison with the reference summaries. First we compared result of our relative entropy based summarizer with all the baseline and top performing systems of DUC 2004, MSE 2005, DUC 2007. The results show that our system performs better than all other systems in all three datasets. We also noticed that when the documents size becomes larger (i.e., the task is more challenging to extract sentences from large diverse collection), our system performs better than the top systems. We then evaluated the performance of different divergence measures as loss functions that can be used in

the proposed framework. Here we notice that all the divergence measures of entropic family perform significantly better than non entropic divergence measures, which signifies the importance of information theoretic and logarithmic relevance in information loss computation. After that, we experimented with different values of redundancy threshold on different datasets. We found that if there is a 40 percent overlap between the sentences already selected as part of summary and a new sentence, then that new sentence should not be a part of summary. Then we experimented with different values of smoothing parameter on different datasets. We noticed that Dirichlet prior smoothing performs slightly better than Jelinek-Mercer smoothing. But the optimal value of smoothing parameter is different for different datasets, and there is no common ideal value for this parameter suiting all the datasets. The performance increase after using smoothing functions is not very significant. This implies that smoothing of parameters does not affect the performance of summarization as much as it affects that of information retrieval. We also noticed that smoothing parameters have greater impact on smaller datasets. After this we compared our work with the previous top system, CLASSY which is a HMM based summarizer and we analyzed the theoretical and practical complexities. We also compared the theoretical complexities of our system with most popular generic summarizer MEAD and other statistical approaches. Finally, in the last section, we analyze and discuss the reason for effective performance of our approach.

Chapter 5

Personalization of Summaries: A New Direction

Different users may have different perspectives on the same text, based on their field of expertise and interest, present summarization systems produce one uniform summary for all users without considering the user's personal interest. Thus there is a great need for summaries to cater to the user's personal background and interests. An effective summarization, thus, should not only be a function of the input text but also of who the reader is and what his prior knowledge is. So a good summary should change in accordance to preferences of its reader. In this chapter we propose to incorporate user inside summarization process. We model user in the proposed information loss based framework of sentence extraction to extract user specific personalized summary for knowledge workers.

5.1 Motivation

One of the issues studied ever since the inception of automatic summarization in the 1960s was that of human agreement[41]: different people can choose different content for their summaries. Marcu-1997[39] found percent agreement of 13 judges over 5 texts from scientific America is 71

percent. Rath-1961[41] found that extracts selected by four different human judges had only 25 percent overlap. Salton-1997[42] found that most important 20 paragraphs extracted by 2 subjects have only 46 percent overlap. These results show that each person has different perspective on the same text and when persons of different background and expertise summarize the same articles, they include different content from each other, reflecting their personal interest and background knowledge. Thus there is a need to incorporate user knowledge in the automatic summarization process to provide them specific summaries. As a real life example consider a speaker who wants to present a summary of his work on a particular topic. The content of his talk would differ based on the background knowledge and expertise level of audience currently present so that the audience can relate to and find it interesting.

At present most summarizers generate summaries using a generic notion of salience. In other words, what is important to summarize is determined by features of the text only, not by who the reader is, or what his background knowledge is. Here we explore the possibility of adding user personal knowledge into automatic summarization process. We treat Summarization process as not only a function of the input text but also of its reader.

5.2 User Specific Summarization

In chapter 3, we proposed information loss framework for document summarization, we treated summarization as a decision making problem and derived a general extraction mechanism for picking sentences based on an ascending order of the expected risk of information loss. We considered each sentence as a possible action and with each such action there is an associated information loss which specifies our decision preferences. But these decision preference may change based on user personal interest. A decision making process may also involves user's own perspective i.e. a person may choose different set of actions based on his knowledge and background. In the summarization setting action space is the entire sentence collection, and if we consider user personal preference the sentence selection process will differ for different users. So the extraction process

which is based on estimation of risk of information loss will differ for different users. We measured this information loss based on a loss function between a sentence and document probability distribution which is independent of any other outside factor. In next section we introduce a user interest factor to influence the information loss computation process.

5.2.1 User Dependent Loss

As discussed in chapter 3, In document summarization setting $\{s_1, s_2..., s_k\}$ are all the possible actions about document set D. There is a risk involve in picking a sentence s_i towards summary, now we consider this risk as user dependent function. So the risk of picking of a sentence depends upon document source D as well as user factor U. The risk associated with a sentence s_i may differ for a user U_x and U_y. So with each action s_i there is associated a loss $L(s_i, D, U)$, which depends upon the document source D as well as the user factor U. To incorporate this user factor in the loss estimation process of summarization we propose to estimate the document model for individual user. We know that every term in document set will have different distribution for different users. For example when a person read some news article, the presence of terms 'cricket' or 'India' will have a different weight for him based on his personal interest and location. So when we estimate a document model the term's distribution should differ with respect the user profile.

To estimate the risk of picking a sentence we need a sentence representation, a user specific document representation, and a loss function to predict document from sentence model. So the risk estimation process is modeled into three basic components: (1) A sentence can be viewed as an observation from a probabilistic sentence model (2) A document set can be viewed as an observation from a probabilistic mixture model of document set and user profile (3) The risk of picking a sentence for a user is a loss between sentence model and user specific document model, that is measured using an intrinsic loss mechanism between the sentence and document distribution.

5.2.2 Document and Sentence Representation

In chapter 3, we represent Document cluster D as a probability distribution of terms. For the document cluster D, we estimate $P(w|D)$,

$$P(w|M_D) = P_{ml}(w|M_D) = \frac{tf(w,D)}{|D|}$$

where $tf(w,D)$ is the frequency of word w in the document D and $|D| = \sum_w D(w)$ is total number of times all words occur in the document set D , it is essentially the length of the document cluster D.

As discussed in the previous section document model will be influenced by presence of user model.

$$P(w|M_{D_U}) = (1-\lambda)P_{ml}(w|M_D) + \lambda P(w|M_U)$$

Here $P(w|M_U)$ is the user model, we will discuss how to estimate the user model in section 5.3. λ is the factor to control the influence of user model.

Sentences are modeled in the same manner as before (section 3.2).

$$P(w|M_S) = \frac{tf(w,S)}{|S|} \tag{5.1}$$

5.2.3 Risk Estimation

We estimate the risk of picking a sentence to be part of summary for a user. Risk of picking a sentence is proportional to loss of information when we consider sentence as a unit to represent the whole document cluster for a particular user. We measure this according to the information theoretic intrinsic loss function relative entropy.

$$Risk(S_i) \propto RE(P(w|M_{S_i}), P(w|M_{D_U})) \tag{5.2}$$

$$= \sum_w P(w|M_{S_i}) log \frac{P(w|M_{S_i})}{P(w|M_{D_U})} \tag{5.3}$$

5.2.4 Estimating User Background model: Experimental Setup

Our aim is to provide specific summaries to user based on their personal interest, and we need user personal data to implement and evaluate our approach. More specifically we need to estimate $P(w|M_U)$. Collecting user's personal data to model his interest has major privacy concerns and other issues. So we proposed a web based profile creation system; where we make use of user's personal data available on web to define their profile. Web is a huge source of information and it contains a lot of personal information of web users. For example for a research professional his information can be in an affiliation page, a project page, a conference page, an online paper, or even in a blog written by himself or others about him. These pages contain enough personal data to model the users. Major benefit here is that there are no privacy issues in user modeling, since the personal data can be obtained anonymously and without any effort from user.

We use search engine to acquire these web pages. It is reasonable to use a search engine because it can search the whole World Wide Web and also tracks the temporal variance of the information available on the Web. For profile creation the first step is to put the person's full name to a search engine (name is quoted with double quotation such as "Albert Einstein") and retrieve documents related to the person. From the search results 'n' top documents are taken and retrieved from their corresponding source websites to define that person's profile. These documents are parsed to extract text content. After performing the removal of stop words and stemming, a unigram language model is learned on the extracted text content. This model can be interpreted as the probability of a word w being related to the person's profile U.

$$P(w|U) = \frac{tf(w,U)}{|U|}$$

With this web based profile creation process we only target online active professional users or knowledge workers but the framework is not restricted only to this domain. This is an experimental setting to evaluate to process of personalized summarization. User studies related to personalization generally suffers from the problem of user personal data and in the university environments this is an easy and effective way to carry out evaluation and experiments related to user modeling.

5.3 Summary Generation

For summary generation, we select sentences with minimum risk to produce summary while keeping the redundancy minimum. The process in similar to summary generation process in section 5.3, where after identification of sentences which are most significant based on their risk value, the redundancy removal step attempts to reduce the size of this set without any reduction in information content. This is achieved by removing the sentences which duplicate information given in other sentences. Once sufficient number of non redundant sentences is picked to make the required length of summary, we arranged them based on chronological ordering (between documents i.e. based on the time stamp) and order of occurrence (within the document). Any additional words than the required length of summary are truncated.

Following is the example (table 5.1) showcasing our technique. The Topic of summary generation is ”Microsoft to open research lab in India”, 8 articles published in different new sources forms the news cluster. A generic summary (using the generic , and User specific summaries for all user were generated from the news cluster, In the example we are showing the condensed summary(100 words) for two users. User A is from NLP domain and User B from network security domain. The italic text in user specific summary shows the differnce compare to generic summary.

5.4 Evaluation

We carried a user based evaluation for personalize summaries. The evaluation of this technique was carried out on five different research scholars working in different fields of computer science. As explained in the previous section web based profile has been built for each of the researchers. News articles of science and technology domain were considered for summarization. Twenty five different topics were chosen with each topic having 5-10 articles. For each topic a generic and user specific summary was generated for each person.

Each researcher was asked to judge the relevance of both versions of summaries for all 25 topics. They have been asked to evaluate the informative ness of summaries on 5 point scale, so

Table 5.1: Example: System Generated Generic and Personalize Summaries

Generic summary: The New Lab, Called Microsoft Research India, Goes Online In January, And Will Be Part Of A Network Of Five Research Labs That Microsoft Runs Worldwide, Said Padmanabhan Anandan, Managing Director Of Microsoft Research India. Microsoft's Mission India, Formally Inaugurated Jan. 12, 2005, Is Microsoft's Third Basic Research Facility Established Outside The United States.In Line With Microsoft's Research Strategy Worldwide,The Bangalore Lab Will Collaborate With And Fund Research At Key Educational Institutions In India, Such As The Indian Institutes Of Technology, Anandan Said.Although Microsoft Research Doesn't Engage In Product Development Itself, Technologies Researchers Create Can Make Their Way Into The Products The Company

User A Specific summary: The New Lab, Called Microsoft Research India, Goes Online In January, And Will Be Part Of A Network Of Five Research Labs That Microsoft Runs Worldwide, Said Padmanabhan Anandan, Managing Director Of Microsoft Research India.Microsoft's Mission India, Formally Inaugurated Jan. 12, 2005, Is Microsoft's Third Basic Research Facility Established Outside The United States. ***Microsoft Will Collaborate With The Government Of India And The Indian Scientific Community To Conduct Research In Indic Language Computing Technologies, This Will Include Areas Such As Machine Translation Between Indian Languages And English, Search And Browsing And Character Recognition.*** In Line With Microsoft's Research Strategy Worldwide,The Bangalore Lab

User B Specific summary: The New Lab, Called Microsoft Research India, Goes Online In January, And Will Be Part Of A Network Of Five Research Labs That Microsoft Runs Worldwide, Said Padmanabhan Anandan, Managing Director Of Microsoft Research India. ***The Newly Announced India Research Group Focuses On Cryptography, Security, Algorithms And Multimedia Security, Ramarathnam Venkatesan, A Leading Cryptographer At Microsoft Research In Redmond, Washington, In The US, Will Head The New Group. Microsoft Research India will conduct a four-week summer school featuring lectures by leading experts in the fields of cryptography, algorithms and security. The program is aimed at senior undergraduate students, graduate students and faculty***

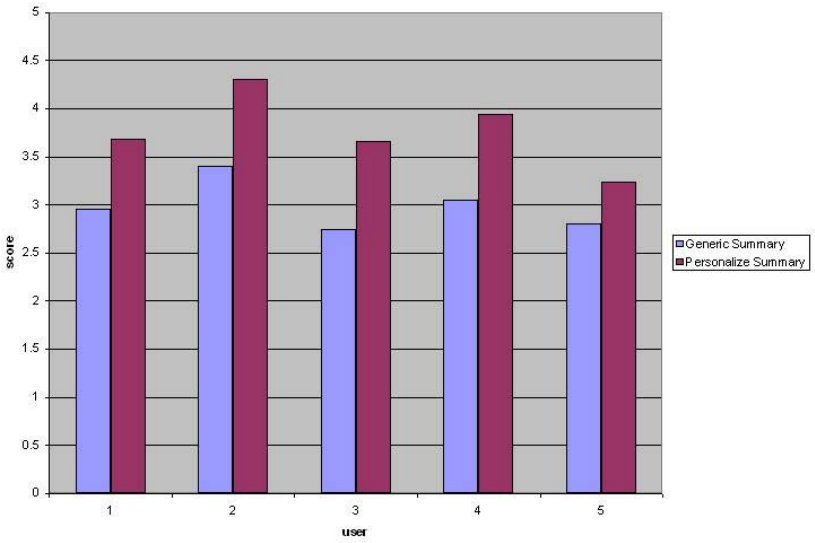

Figure 5.1: Average Scores for different Users

each user provide manual scores to both generic and personalize summaries based on how relevant the summaries are for them. The average scores for both types of summaries of all topics for each researcher is shown in Figure 5.1. It shows that the users prefer profile based personalized summaries compared to a generic summary given by general automatic summarization system. This means that personalization can benefit the automatic summarization process to improve user satisfaction towards summaries.

Figure 5.2 shows the scores given by a particular user across different topics. We see that for most of the topics user find personalized summaries relevant for him. Also the personalized summaries for the topics strongly related to the user's domain are more relevant to him. For topics which are not closely related to user's field, the personalized and generic summaries are quite similar. These topics are the ones which got least influenced with personalization. For a few rare topics the user did not find personalized summary better, which may be because of presence of noisy data in their profile.

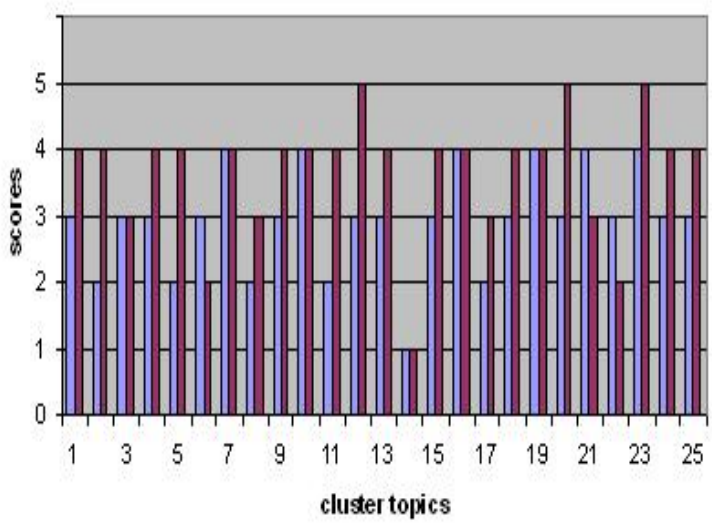

Figure 5.2: Score of Different topics for a User

5.5 Summary

In this chapter, we go beyond the traditional notion of generic relevance and incorporate user personal interest in document summarization process. Here we treat summarization process as not only a function of the input text but also of its reader. We first discussed the need and motivation to generate personalized summaries. We were mainly motivated from the fact that each person has different perspective on the same text and when persons of different background and expertise summarize the same articles, they include different content, reflecting their personal interest and background knowledge. We extended our information loss framework to include user into the sentence extraction framework. Here risk of picking a sentence depends upon document source as well as user factor. We incorporated user factor in the document model estimation, as we know that every term in document set will have different distribution for different users. To estimate user influence, we used their personal data available on web to create their profile and estimated the user model. To evaluate personalized summaries, a controlled user-centered qualitative evaluation was carried out on news articles of science and technology domain. Each user was asked to judge the

relevance of both versions of summaries for all topics. The results indicated better user satisfaction with personalized summaries compared to generic summaries.

Chapter 6

Conclusion and Future Work

This thesis presents a new framework for extractive document summarization based on information loss. In this framework, sentences and documents are modeled using probabilistic language models and a loss function is used to estimate the relevance of sentence to be picked as part of summary.

We have explored several loss functions in the framework. Use of information theoretic relative entropy risk function is most natural and effective. It measures the amount of information loss when a sentence has been chosen instead of whole document cluster. Sentences with minimum information loss are considered to be most important. The idea is to measures how bad a sentence distribution is in modeling the documents distribution. It presents a unifying view of summarization, as method to approximate a large document set by a smaller summary with minimum information loss. We approximate a document set by a small sentence set. The candidate of summary act as a surrogate for document set in a larger inference process.

One fundamental difference between the Information loss framework and any existing extraction mechanism is that the information loss framework treats the summarization problem as a decision problem, and incorporates probabilistic language models and loss function as major components in the framework. While previous work uses combination of various features to calculate sentence

importance we use information loss as a single measure to evaluate the worthiness of sentence to be part of summary.

There are other factors in multi document summarization such as redundancy removal and readability, and we use simplistic methods like terms overlap and chronological ordering for this purpose. Through the proposed sentence selection, along with a simple redundancy measure and text reformulation mechanism, we come up with a light-weight summarizer to generate more informative summary than the earlier approaches which use very complex algorithm for summary generation. Our algorithm generates extracts on the fly without extensive computation or training or the estimation and parameterization of multiple features which seems to be used in various state of the art algorithms.

In order to evaluate the performance of our approach, we used different data sets that have been used in recent document summarization: DUC (Document Understanding Conference) competition different year datasets and MSE (Multi-Lingual Summarization Evaluation) dataset. We used the automatic summary evaluation metric, ROUGE which is the standard way of evaluation of summaries. The evaluation results show that our approach outperforms all the reported systems for all three metrics ROUGE-1 ROUGE-2 and ROUGE-SU4.

We studied the effect of different values of redundancy threshold on summarization performance and found that a 40 percent term overlap between sentences is a good heuristic to estimate redundancy. Studying the effect of smoothing probabilities of language models is another contribution of this thesis. We evaluated popular smoothing methods (Jelinek-Mercer, Dirichlet priors), and found that the summarization performance is not that sensitive to the smoothing parameters as compare to its impact on information retrieval systems.

This thesis also explores the possibility of incorporating user specific content in summary gen-

eration. We go beyond the traditional notion of generic relevance and incorporate user factor as sentence extraction criteria. We believe that a good summary should change in accordance to preferences of its reader. For this purpose we model user in the proposed information loss based framework of sentence extraction to extract user specific personalized summaries. We create a web based profile for knowledge workers using their personal data available on web to model their background and interests. To evaluate personalized summaries, a controlled user-centered qualitative evaluation was carried out on news articles of science and technology domain. The results indicate better user satisfaction with personalized summaries compared to generic summaries.

6.1 Future Directions

The information loss framework provides a general probabilistic framework including the language modeling approach. This provides a connection between text summarization and statistical language models. It opens up new possibilities for developing principled approaches to text summarization as it allows exploring different extraction models systematically through considering different loss functions and using different language models in the framework.

In the process of sentence extraction during summary generation an interactive summarization interface can be considered to enhance to estimate a better sentence model. In interactive summarization scenario interface user has the option to provide relevance feedback by choosing a set of sentences as relevant to be part of summary. The proposed framework explicitly deals with the sentence model and can perform feedback more naturally by treating it as sentence model updating, that can eventually improve the quality of summary.

Present work mainly concentrates to present an effective content selection process. More sophisticated natural languages processing approaches like entity dereferencing and co reference resolution can be applied for better coherence after sentence extraction in order to improve the readability.

The generation of human like abstractive summaries is complex task; however there have been

efforts to generate some sort of abstractive summaries which use text analysis and processing techniques on top of extracts. Our lightweight approach to generate extracts provides a very good base for such approaches to generate abstractive summaries.

The proposed design of generating personalized summaries is not restricted to web based profile creation for a user. Current results with the proposed user model are encouraging and this motivates to carry out these experiments with other richer ways of building user background models to benefit more users and community in future.

Bibliography

[1] Cover, T. M., and Thomas, J. A. 1991 Elements of Information Theory. Wiley-Interscience, New York, New York, 1991.

[2] Lindgren, B.W. 1971 Elements of Decision Theory. New York: The Macmillan Company.

[3] A. Rnyi 1961. ”On measures of information and entropy”. Proceedings of the 4th Berkeley Symposium on Mathematics, Statistics and Probability 1960: 547-561.

[4] Robert, C. P. 1996. Intrinsic losses. Theory and Decision 40, 191214.

[5] David Haussler, Manfred Opper: Metric Entropy and Minimax Risk in Classification. Structures in Logic and Computer Science 1997: 212-235

[6] H. P. Luhn. The Automatic Creation of Literature Abstracts. In IBM Journal of Research and Development, pages 159-165, 1958.

[7] C.Y. Lin and E.H. Hovy 2003 Automatic Evaluation of Summaries Using N-gram Co-occurrence Statistics. In Proceedings of HLT-NAACL 2003

[8] Inderjeet Mani. Automatic summarization. Computational Linguistics - MIT Press, 2001.

[9] Inderjeet Mani and Mark Maybury. Advances in Automatic Text Summarization. MIT Press, 1999.

[10] J. Berger 1985 Statistical decision theory and Bayesian analysis, New York: Springer-Verlag New York Inc.

[11] Bather, J. 2000. Decision Theory. Chichester: John Wiley and Sons, Inc.

[12] P. Over and J. Yen. 2004 An introduction to DUC 2004 intrinsic evaluation of generic news text summarization systems. In Proceedings of DUC, 2004.

[13] D. R. Radev, H. Y. Jing, M. Stys and D. Tam 2004 Centroid-based summarization of multiple documents. Information Processing and Management, 40: 919-938, 2004

[14] I. Mani and E. Bloedorn 2000 Summarizing Similarities and Differences Among Related Documents. Journal of Information Retrieval,2000.

[15] U. Hahn and I. Mani. The challenges of automatic summarization. IEEE Computer, 33(11):2935, 2000.

[16] I. Mani, D. House, G. Klein, L. Hirschman, L. Obrst, T. Firmin, M. Chrzanowski, and B. Sundheim. The TIPSTER SUMMAC Text Summarization Evaluation: Final report. In Technical report, DARPA, 1998.

[17] H. P. Edmundson. New methods in automatic extraction. Journal of the ACM, 16(2):264285, 1969.

[18] L Page, S Brin, R Motwani, T Winograd. The pagerank citation ranking: Bringing order to the web, Technical Report, Standford Digital Library Technologies Project 1998.

[19] Jelinek, F. (1997). Statistical methods for speech recognition. MIT Press. 1997.

[20] Rosenfeld, R. (2000). Two decades of statistical language modeling: where do we go from here? In Proceedings of IEEE, volume 88. 2000.

[21] R. Mihalcea and P. Tarau. 2004. TextRank bringing order into text, EMNLP 2004, 404-411, 2004.

[22] G. Erkan and D. Radev 2004 LexPageRank: prestige in multidocument text summarization. in Proceedings of EMNLP 2004

[23] Gunes Erkan and Dragomir Radev. 2004. LexRank: Graph-based Centrality as Salience in Text Summarization. Journal of Artificial Intelligence Research, 2004.

[24] V. Lavrenko and W. B. Croft 2001 Relevance based languagemodels. In SIGIR 01: Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval

[25] J. Lafferty and C. Zhai 2001 Probabilistic IR models based on document and query generation. Proceedings of the workshop on Language Modeling and Information Retrieval

[26] J. Conroy, J. Schlesinger, J. Goldstein, and D. OLeary. 2004 Left-brain/right-brain multi-document summarization. In Proceedings of DUC, 2004.

[27] J. Conroy, J. Schlesinger, and J. Goldstein. 2005 Three classy ways to perform arabic and english multidocument summarization. In Proc. of MSE, 2005

[28] S. Harabagiu and F. Lacatusu 2005 Topic themes for multidocument summarization. In Proceedings of SIGIR, Salvador, Brazil, 202-209, 2005

[29] C. Lin and E. Hovy. The automatic acquistion of topic signatures for text summarization. In Proc. of COLING, 2000

[30] C.Y. Lin and E.H. Hovy 2002 From Single to Multidocument Summarization: A Prototype System and its Evaluation. In Proceedings of ACL-2002.

[31] Daume H., and Marcu D 2006 Bayesian query-focused summarization. In Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (pp.05-312).

[32] W.T. Yih, J. Goodman, L. Vanderwende, and H. Suzuki. 2007 Multi-document summarization by maximizing informative content words. In IJCAI 2007: 20th International Joint Conference on Artificial Intelligence, January, 2007.

[33] Manning, C., and Schutze, H. 1999 Foundations of Statistical Natural Language Processing. The MIT Press, 1999.

[34] Kupiec J., Pederson J., Chen F.A. 1995 Trainable Document Summarizer. Proceedings of the 18th ACM SIGIR, 68-73, 1995.

[35] Amini M.-R., Gallinari P. 2002 The Use of unlabeled data to improve supervised learning for text summarization. Proceedings of the 25th ACM SIGIR, 105-112, 2002.

[36] Hardy, H., Shimizu, N., Strzalkowski, T., Ting, L., Wise, G. B., Zhang, X. 2002 Cross-document summarization by concept classification. In Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Tampere, Finland.

[37] A. Nenkova and L. Vanderwende.:The impact of frequency on summarization. Technical report, MSR-TR-2005-101, 2005.

[38] A. Nenkova, L. Vanderwende, and K. McKeown.: A compositional context sensitive multi-document summarizer. In Proc. of SIGIR, 2006.

[39] Daniel Marcu. From Discourse Structures to Text Summaries. Proceedings of the 14th National Conference on Artificial Intelligence AAAI-97

[40] Zhai and J. Lafferty. A study of smoothing methods for language models applied to information retrieval. ACM Trans. Inf. Syst., 22(2):179214, 2004

[41] GJ Rath, A Resnick, TR Savage. The formation of abstracts by the selection of sentences. American Documentation,12(2): 139143, April 1961.

[42] G Salton, A Singhal, M Mitra, C Buckley. Automatic text structuring and summarization. Information Processing and Management,33(2): 193-207, 1997.

[43] Document Understanding Conference.http://wwwnlpir. nist.gov/projects/duc/

[44] C. E. Shanon and W. Weaver.: Mathematical Theory of Communication. University of Illinois Press. 1983.

[45] L. Douglas Baker and Andrew Kachites McCallum.: Distributional Clustering of Words for Text Classification. In SIGIR 98: pages 96103, New York, NY, USA. ACM Press. 1998.

[46] Jade Goldstein, Mark Kantrowitz, Vibhu O. Mittal, and Jaime Carbonell. Summarizing Text Documents:Sentence Selection and Evaluation Metrics. In Proceedings of SIGIR-99, Berkeley, CA, August 1999.

[47] M. Jaoua and A. Ben Hamadou.: Automatic text summarization of scientific articles based on classification of extract's population. In Proceedings of Computational Linguistics and Intelligent Text Processing, 2003.

[48] E. Hellinger, Neue Begrundung der Theorie der quadratischen Formen von unendlichen vielen Veranderlichen, J. Reine Aug. Math., 136(1909), 210-271.

[49] S. S. Dragmir, J. Sunde and C. Buse, New Inequalities for Jeffreys Divergence Measure, Tamsui Oxford Journal of Mathematical Sciences, 16(2)(2000), 295-309.

[50] H. Jeffrays, An Invariant Form for the Prior Probability in Estimation Problems, Proc. Roy. Soc. Lon., Ser. A, 186(1946), 453-461.

[51] R. Sibson, Information Radius, Z. Wahrs. und verw Geb., 14(1969), 149-160.

[52] J. Burbea, J. and C.R. Rao, Entropy Differential Metric, Distance and Divergence Measures in Probability Spaces: A Unified Approach, J. Multi. Analysis, 12(1982), 575-596.

[53] J. Burbea, J. and C.R. Rao, On the Convexity of Some Divergence Measures Based on Entropy Functions, IEEE Trans. on Inform. Theory, IT-28(1982), 489-495.

[54] K. Pearson, On the Criterion that a given system of eviations from the probable in the case of correlated system of variables is such that it can be reasonable supposed to have arisen from random sampling, Phil. Mag., 50(1900), 157-172.

[55] I. J. Taneza, New Developments in Generalized Information Measures, Chapter in: Advances in Imaging and Electron Physics, Ed. P.W. Hawkes, 91(1995), 37-136.

[56] Lin J., Wong S. K. M. A new directed divergence measure and its characterization, Int. J. General Systems, 17:7381,. 1990

[57] Bernardo, J. M. Expected information as expected utility. Ann. Statist. 7, 686690, 1979.

[58] Bernardo, J. M. and Smith, A. F. M. . Bayesian Theory. Chichester: Wiley 1994.

[59] Kathleen R. McKeown, Regina Barzilay, David Evans, Vasileios Hatzivassiloglou, Judith L. Klavans, Ani Nenkova, Carl Sable, Barry Schiffman and Sergey Sigelman Tracking and Summarizing News on a Daily Basis with Columbia's Newsblaster in Proceedings of HLT 2002 Human Language Technology Conference, San Diego, CA, 2002.

[60] Kathleen McKeown, Regina Barzilay, John Chen, David Elson, David Evans, Judith Klavans, Ani Nenkova, Barry Schiffman and Sergey Sigelman, Columbia's Newsblaster: New Features and Future Directions Demo at NAACL-HLT 2003.

[61] Kathleen McKeown, Rebecca Passonneau, David Elson, Ani Nenkova, Julia Hirschberg, Do Summaries Help? A Task-Based Evaluation of Multi-Document Summarization 28th Annual International ACM SIGIR Conference on Research and Development in Information Retreival, Salvador, Brazil, 2005.

[62] J. Galliers and K. Jones. Evaluating Natural Language Processing Systems. In Lecture Notes in Artificial Intelligence. Springer, 1995.

[63] J. Minel, S. Nugier, and P. Gerald. How to Appreciate the Quality of Automatic Text Summarization. In Proc. of a Workshop, ACL, pages 25-30, 1997.

[64] D. Radev, H. Jing, and M. Budzikowska. Centroid-based summarization of multiple documents: sentence extraction, Utility-Based Evaluation, and User Studies. In ANLP-NAACL 2000 Workshop, pages 21-29, 2000.

[65] Jean Carletta. Assessing agreement on classification tasks: The kappa statistic. Computational Linguistics, 22(2):249-254, 1996.

[66] K. Papineni, S. Roukos, T. Ward, and W. Zhu. BLEU: A Method for Automatic Evaluation of Machine Translation, 2001.

[67] H. Jing, R. Barzilay, K. McKeown, and M. Elhadad. Summarization Evaluation Methods: Experiments and Analysis. In AAAI Intelligent Text Summarization Workshop, pages 60-68, 1998.