Helmholtz Principle-Based Keyword Extraction

Anima Pradhan

Department of Computer Science and Engineering
National Institute of Technology Rourkela
Rourkela-769 008, Odisha, India

Helmholtz Principle-Based Keyword Extraction

Thesis submitted in

May 2013

to the department of

Computer Science and Engineering

of

National Institute of Technology Rourkela

in partial fulfillment of the requirements

for the degree of

Master of Technology

in

Computer Science and Engineering

Specialization : Computer Science

by

Anima Pradhan

[Roll No. 211cs1048]

under the guidance of

Asst. Prof. Korra Sathya Babu

Department of Computer Science and Engineering
National Institute of Technology Rourkela
Rourkela-769 008, Odisha, India

Department of Computer Science & Engineering
National Institute of Technology Rourkela
Rourkela-769 008, Odisha, India. www.nitrkl.ac.in

Korra Sathya Babu
Asst. Professor

May 20, 2011

Certificate

This is to certify that the work in the thesis entitled ***Helmholtz Principle-Based Keyword Extraction*** **by** ***Anima Pradhan*****, bearing Roll No. 211CS1048, is a record of an original research work carried out by him under my supervision and guidance in partial fulfilment of the requirements for the award of the degree of** *Master of Technology* **in** *Computer Science and Engineering***. Neither this thesis nor any part of it has been submitted for any degree or academic award elsewhere.**

Korra Sathya Babu

Department of Computer Science and Engineering
National Institute of Technology Rourkela
Rourkela-769 008, India. `www.nitrkl.ac.in`

Mr. Korra Sathya Babu
Assistant Professor

May, 2013

Certificate

This is to certify that the work in the project entitled *Keyword Extraction Based Helmholtz Principle* **by** *Anima Pradhan* **is a record of an original work carried out by her under my supervision and guidance in partial fulfillment of the requirements for the award of the degree of** ***Master of Technology*** **in** ***Computer Science and Engineering*****. Neither this project nor any part of it has been submitted for any degree or academic award elsewhere.**

Korra Sathya Babu

Acknowledgment

I would like to express my gratitude to my thesis guide Asst. Prof. Korra Sathya Babu for the useful comments, remarks and engagement through the learning process of this master thesis. The flexibility of work he has offered me has deeply encouraged me producing the research.

Furthermore I would like to thank Dr. Pankaj Kumar Sa for for being a source of support and motivation for carrying out quality work.

My hearty thanks goes to Mr. Sambit Bakshi for consistently showing me innovative research directions for the entire period of carrying out the research and helping me in shaping up the thesis. Also, I like to thank the participants in my survey, who have willingly shared their precious time during the process of interviewing. I would like to thank my loved ones, who have supported me throughout entire process, both by keeping me harmonious and helping me putting pieces together. I will be grateful forever for your love.

Last but not the least,i like to thank my family and the one above all of us, the omnipresent God, for answering my prayers for giving me the strength to plod on despite my constitution wanting to give up, thank you so much Dear Lord.

Anima Pradhan

Abstract

In today's world of evolving technology, everybody wishes to accomplish tasks in least time. As information available online is perpetuating every day, it becomes very difficult to summarize any more than 100 documents in acceptable time. Thus, "text summarization" is a challenging problem in the area of Natural Language Processing (NLP) especially in the context of global languages.

In this thesis, we survey taxonomy of text summarization from different aspects. It briefly explains different approaches to summarization and the evaluation parameters. Also presented are a thorough details and facts about more than fifty automatic text summarization systems to ease the job of researchers and serve as a short encyclopedia for the investigated systems.

Keyword extraction methods plays vital role in text mining and document processing. Keywords represent essential content of a document. Text mining applications take the advantage of keywords for processing documents. A quality Keyword is a word that represents the exact content of the text subsetly. It is very difficult to process large number of documents to get high quality keywords in acceptable time.

This thesis gives a comparison between the most popular keyword extractions method, tf-idf and the proposed method that is based on Helmholtz Principle. Helmholtz Principle is based on the ideas from image processing and derived from the Gestalt theory of human perception. We also investigate the run time to extract the keywords by both the methods. Experimental results show that keyword extraction method based on Helmholtz Principle outperformancetf-idf.

Keywords: **Text Mining, Text Summarization, Stemming, Helmholtz Peinciple, Information Retrieval, Keyword Extraction, Term Frequency - Inverse Document Frequency.**

Contents

List of Algorithms

List of Figures

List of Tables

Chapter 1

Introduction

In this thesis We have done briefly survey on Automatic Text Summarizers which help us to have an idea of what Text Summarization is and how it can be useful for. Also We propose approaches for comparison of keyword extraction using term weighting and Helmholtz Principle in multi documents. We focus on two text mining tasks: text summarization and keyword extraction. We aim to identify and tackle the challenges of multi documents and compare the performance of the proposed approaches against a wide range of existing methods. Text mining, sometimes alternately referred to as text data mining, roughly equivalent to text analytics, refers to the process of deriving high-quality information from text. It is a well research field; for instance, during the 1990's and early 2000 text summarization received a lot of attention due to its relevance to both information retrieval and machine learning.

There are several approaches to term weighting of which the Term Frequency - Inverse Document Frequency (TF-IDF) is probably the most often used. It is an approach that relies heavily on term frequency (TF); i.e., a statistic of how many times a word appears within a document. In many cases, TF is a good statistic to measure the importance of a word: if it occurs often, it could be important.

1.1 Text Summarization

A summary is a reduced transformation from original text through selection and generalization of the important concept [2]. Summarization model consists of three

stages:

- **Interpretation** : **Original text converted into structured representation so that necessary computation and modification can be performed on it.**
- **Transformation** : **convert into summary representation.**
- **Generalization** : **summary representation converted into summary text.**

Effective summarizing requires an explicit, and detailed, analysis of context factors, as is apparent when we recognize that what summaries should be like is defined by what they are wanted for, as well as by what their sources are like [3]. Context Factor distinguishes three main factors:

1.1.1 Input Factor

The features of input document can affect the resulting summary according to the following aspects :

Document Structure

Structure is a explicit organization of a Input Document. Examples are : **header, chapters, sections, lists, table etc. Structure of the Document should be well organized, so that information can be use to analyze the document.**
Summarizer [4], PALSUMM that create summaries by choosing sentences or parts of sentences corresponding to nodes at a given level of depth of a tree structured representation of the structure of the text produce excellent summaries of the original text [5]shows structural properties of medical articles.

Domain

Domainsensitive systems are able to obtain summaries of single or specific topic domain (e.g. all of medicine as a single domain) with varying degrees of probability. For example [6] applied two independent method (BIOChain and FreqDist) for identifying salient sentences in biomedical texts. [7] shows how argumentation schemes and story schemes form most relevant forms of commonsense knowledge

in the context of reasoning with evidence. Some other domain specific information summarizer for different kinds of documents.

Specialization level

A text may be broadly characterized as ordinary, specialized, or restricted, in relation to the presumed subject knowledge of the source text readers. This aspect can be considered same as domain aspect.

Language

The language of the input can be general language or restricted to sub language within a domain, purpose or audience. Summarization algorithm may or may not use language dependent information. Considering specific form factors, TIDES include information detection, extraction, summarization and translation focusing currently on English, Chinese and Arabic with some research on Korean and Spanish. LDC work on Chinese and Arabic language. English has been the main language (see DUC), with substantial effort in Japanese (see NTCIR) and work on Chinese and German, and both raw Arabic and automaticallytranslated Arabic news in DUC .

Media

Although Our main focus of Summarization is textual summarization but summaries of nontextual documents like , audio, video [8],Multimedia, Images etc.Summarizing of Multimedia resources by the technology DREL [9]. To achieve consistency of image content representation and highquality results, imagebased summarization needs to be geared toward specific image types [10].

Unit

Different Number documents can be used to create summary of the document. If only single document is used to create summary, it is named as Single Document Summarization System. If more than one document is used, then it is named as MultiDocument Summarization System. In Multidocument Summarization system does not simply shorten the source texts but presents information organized around

the key aspects to represent a wider diversity views on the topic. Different types of Summarizer for different kinds of documents developed by Columbia University are: SUMMONS, MultiGen, FociSum.University of South California produces a summarizer system, Summarist.SUMMONS and MultiGen works for news domain where FociSum based on question and answer approach. Summarist produces summaries of Web Documents.

Genre

Some systems exploit typical genredetermined characteristics of texts, like pyramidal organization of newspaper article ,development of scientific article, etc. Some summarizers are independent of type of documents but some are specialized on some certain type of documents like Broadcast fragments [11], e-mails [12][13],web pages [14], news, medical articles [15], scientific articles[16], News agency [17]etc.

Scale

Scale means length of Input source. Length of input documents can be varies. Longer documents like reports, books contains more important informative parts, contain more topics, less redundant information, etc. Where shorter document like news articles, sentences contain less information, contain less topic, less meaningful information.

1.1.2 Purpose Factors

Here it is describe for what purpose we are doing summarizing. Purpose factors are fall under three categories :

Situation

Situation is the context of Summary. It refers to the environment where summary is to be used. The environment of Summary means, by whom, for what purpose and when it will be used, it may or may not be known. If it is known in advance then it can fulfill

the requirements of context of the summary. For example, Medical literatures on the web are the important sources to help clinicians in patient care.

Audience

Audience refers to the readers for whom summarization is to be done. It may be done according to the interest of the audience.

Use

Use refers to for what reason summarization is to be done. Summaries can be used for retrieving information, developed Search Engine, informationcovering substitutes for their source text, as devices for refreshing the memory of an alreadyread source.

1.1.3 Output Factors

There are at least three major output factors are:

Material

The summary of a document can contain all important concepts of original document or only some aspects of it. Summaries may be designed to contain some specific type of information like, in papers what was observed, plot, etc. Generic summaries cover all important concepts where querybased summaries cover related to the need of user.

Format

created summary organized into different sections like headings, etc. In some journal papers, like an abstracts, or Test results.

Style

A Summary can be :

1. **Informative** : **It cover concept of original document.**
2. **Indicative** : **It gives brief explanation of original document.**

3. **Aggregative : It gives partial information of which does not cover in original document.**

4. **Critical : It review summaries whether it is wrong, right or require some modification.**

1.2 Keyword Extraction

Keyword extraction is highly related to automated text summarization. In text summarization, most indicative sentences are extracted to represent the text. In order to utilize the information from short documents, whether we want to categorize the text or extract information from it, we need to identify which words are the most important within the text. This can be achieved by various methods. I focused on comparison of performance of two keyword extraction methods on very large data set such as very popular method term weighting method TF-IDF and another is based on Helmholtz Principle.

Helmholtz Principle is based on the ideas from image processing and especially on the Helmholtz Principle from the Gestalt Theory of human perception. According to a basic principle of perception due to Helmholtz, an observed geometric structure is perceptually meaningful if it has a very low probability to appear in noise.

1.2.1 Application

Automatic text summarization can be used:

- **To summarize news to SMS or WAP-format for mobile phones/PDA.**
- **To let a computer synthetical read the summarized text. Written text can be to long and boring to listen to.**
- **In search engines to present compressed descriptions of the search results (see the Internet search engine Google).**
- **To search in foreign languages and obtain an automatically translated summary of the automatically summarized text.**

In this chapter, motivation of the research and the outline of the work is introduced.

1.2.2 Motivation

Due to growth of online information it is difficult for human beings to accomplish their task in the field of natural language processing in stipulated time. Huge number of available documents in digital media makes it difficult to obtain the necessary information related to the needs of a user. In order to solve this issue, text summarization systems can be used. The text summarization systems extract brief information from a given document while preserving important concepts of that document. By using the summary produced, a user can decide if a document is related to his/her needs without reading the whole document. Also other systems, such as search engines, news portals etc., can use document summaries to perform their jobs more efficiently.

To extract important information or sentences, high quality keyword plays crucial role as per user requirement. They help users to search information more efficiently. Due to growth of online information it is difficult for human beings to accomplish their task in the field of natural language processing in stipulated time. Extracting high quality keywords automatically are expensive and time consuming. This shows keyword extraction is challenging problem in the area of natural language processing especially in the context of global languages in acceptable time.

Annotation of keyword of document can be used to build keyword query. In an electronic magazine keyword give a clue about the main idea of an article . In a book they quickly lead the leader to the whereabout of the information sought.On the Web, tag annotations help to find multimedia and other resources.Moreover, creation of annotations is time consuming, such that automatic ways of keyword extraction form the document are required.

There are many existing algorithms have been proposed for Automatic Keyword extraction. Helmholtz Principle is developed for mining textual, unstructured or sequential data. Here We define a new measure of meaningful keywords with good performance on different type of documents. TF-IDF is successful and most well

tested technique in Information Retrieval. So we compare most popular method TF-IDF with my proposed algorithm based on Helmholtz Principle for large number of documents.

1.2.3 Thesis Outline

The rest of the thesis is organized as follows:

Chapter 2 presents the related work in documents summarization and keyword extraction methods. Taxonomy of text summarization systems, text summarization approaches in literature, and evaluation measures of the text summarization systems are explained briefly. Approaches of keyword extraction methods are presented.

Chapter 3 explains briefly on automatic text summarizer systems with their features. Also explain TF-IDF method and Helmholtz principle based keyword extraction. I presented proposed algorithm for keyword extraction.

Chapter 4 I present the experimental results.

Chapter 5 presents the concluding remarks.

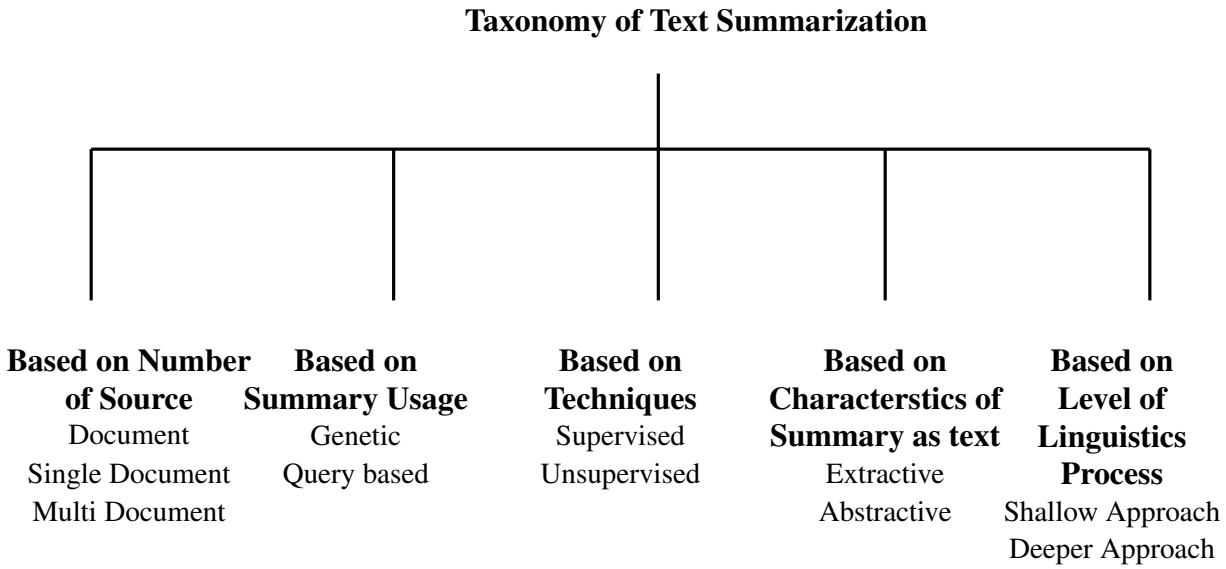
Chapter 2

Related Work

2.1 Taxonomy of Text Summarization

The summary can have different categorization according to their characteristics.

- **Based on Number of source documents : If single document is used for summarization, it is known as SingleDocument Summarization. More than one document is used, and then it is known as Multidocument Summarization.**
- **Based on Summary Usage**
  - **Generic Summarization :If whole document is used for creating summary.**
  - **Query based Summarization : If specific topic is used related to the query.**
- **Based on techniques**
  - **Supervised Summarization : The training data set is known.**
  - **Unsupervised Summarization : Training data set is not known.**
- **Based on characteristics of a summary as text**
  - **Extractive : Its process is to find more important information or sentences from input document to create a summary.**
  - **Abstractive : In this process, machine need to understand the concept of all the input documents then produce summary with its own sentences.**



- **Based on the level in the linguistic space**
  - **Shallow approaches** : **It related to the syntactic level of representation.**
  - **Deeper approach** : **It is related to the semantic level of representation and allows linguistic process at some level.**

2.1.1 Extractive Summarization Method

It finds more important information or sentences from input document to create a summary. There is different level processing to get more informative parts or high concepts information form input document. Based on these levels of processing, text summarization is categorized into different approaches.

2.1.2 Approaches of Text Summarization

Statistical Approaches

In 1958,[18] describe that a sentence gives useful measurement of significance, if frequency of particular term(or word) is high in an article. Term Frequency : **Number of occurrence of words. At the time of implementation, he proposed some key ideas:**

- **Stemming** : **In a document some words can be seen in different variant like**

singular vs Plural, present verses past, past verses future, written in small or capital letter etc. Ex. School, schools, School, SCHOOL all are same. Stemmer is a tool which reduces a word to its root form. For example, reads, reading, read is stemmed into read. Here frequency of read is 3.Advantage is, it reduces the memory usage for storing words. Another stemmer is, Porter Stemmer (Porter Stemmer, 2000) for English document. In 1996, Rao et al. [19] and in 2012, U.Mishra et al. propropose a stemmer MAULIK [20] for Hindi document and in 1999, Zemberek proposed Zemberek Morphological Analyzer for Turkish document.

- **Stop word Removal** : **The words which do not conveying any significance semantic to the text. These are "the","a","an", "from","to","of", etc. Stop word removal is done using human made list of words. This list is different for different languages. Here author applied this scheme in a set of 50 articles. The sentence which comprises more significance words set as highest ranking sentence and keeps all sentences in a decreasing order based on their rank. Then it extracts sentences whose rank is more than predefined threshold value. In 1969, Edmundson [21] introduce four basic methods for automatic extracting system was based on assigning to text sentences numerical weights that were functions of the weights assigned to certain machinerecognizable characteristics. These four basic methods are:**

  1. **Cue Method** : **Relevance of a sentence is affected by presence of pragmatic words ("significant", "impossible" and "hardly"). In this method,Cue dictionary comprises three sub dictionaries** :
     Bonus words : **positively relevant,**
     Stigma words : **Negatively relevant,**
     Null words : **Irrelevant.**

  2. **Key Method** : **According to this, more frequent words are positively relevant. First it finds the total number of word occurrences in the document. The words are set according to the nondecreasing order and the word whose frequencies above the threshold were assumed as Key words**

and assigned positive weights equal to the frequency of occurrence.The final Key weight of a sentence is the sum of the Key weights of its constituent words.

3. **Title Method** : **In this method, the machine recognizes certain specific characteristics of the document, like title, headings, and format. The Title method compiles for each document, a Title glossary comprises of nonNull words of the title, subtitle, and headings for that document. Words in the title glossary are assigned positive weights. The final weight for each sentence is the sum of the Title weights of its constituent words.**

4. **Location Method** : **In the Location method, the sentences which contain specific headings are positively relevant sentences. It selects headings of documents which are appear in corpus and stored in a Heading dictionary. Mostly heading words are appearing in the "Introduction", "Purpose", and "Conclusion" parts of a document. The final Location weight for each sentence is the sum of heading weight.**

Author applied these methods in a set of 400 documents, and find that, the CueTitleLocation method gives highest mean co selection score while Keymethod give less.Emundsons settled features for extracting sentence. These are:

- **Sentence Length Cutoff Feature** : **If a sentence is longer than the pre specified threshold value is more important than shorter sentence.**

- **FixedPhrase Feature** : **If Sentence containing any fixed phrases like "this letter","In conclusion" or following immediately after heading containing a keywords like "conclusion","results","summary" are more important.**

- **Paragraph Feature** : **If a paragraph containing more than one sentence than importance of sentence is based on position, whether it is paragraphinitial or paragraphfinal or paragraphmedial. Paragraphinitial sentence is, more important than Paragraphfinal sentence.**

- **Thematic Word Feature: The most frequent content words is known as thematic words. A sentence is scored based on function of frequency.**

- **Uppercase Word Feature : Proper names are important. e.g. " ASTM (American Society for Testing Materials)".This feature is computed with the constraints that an uppercase thematic word is not sentenceinitial and begin with a capital letter. Actions are : TFIDF, entropy, mutual information and statistics. Another Statistical approaches used for keyword extraction are : TFIDF, entropy, mutual information and statistics [22],[23].**

Coherent Based Approach

A coherent based approach basically deals with the cohesion relations among the words.Cohesion relations among elements in a text: reference, ellipsis, substitution, conjuction, and lexical cohesion [24].

- **Lexical chain : Lexical chain is a method of identifying set of words which are semantically related. Semantic relationships among the words can be systematic semantic, and nonsystematic semantic.**
  Semantically related words can be extracted using dictionaries and WordNet.

- **WordNet : In NLP WordNet is used for measuring of conceptually similarity and relatedness information from document. Concept can be related in any ways beyond similar to each other. For Example, a wheel is a part of a car, night is the opposite of day and so forth [25],[26].**
  In [27] describe four features based on lexical chains. Features are:

- **Lexical chain score of a word : A word can be a member of more than one lexical chains as it can appear in a same text with different sense. The score of a lexical chain depends on relations appearing in the lexical chain.**

- **Direct Lexical chain score of a word : Score was calculated based on the relations that belong to the word.**

- **Lexical chain span score of a word : It depends on the portion of the text that is covered by the lexical chain. The covered portion of the text is considered as the**

distance between the first positions of a lexical chain member (word) occurred first in the text and the last occurrence position of a lexical chain member which occurred last in the text.

- **Direct lexical chain span score of a word** : **It is computed same as the lexical chain span score except that it is considered the words which are directly related with the word in the lexical chain.Author applied these four features with a corpus consists of 155 abstracts and got 45% precision in the extraction of keywords. In [27] propose a CRF based keyword extraction approach.Rhetorical Structure Theory (RST) based methods are another example that uses Coherent based summarization.**

- **RST**: **It organizes texts into treelike structure to represent the coherence relations among the words [28]. In [29] propose a automatic summarizer GIST based on RST processes. In [30] propose a Automatic text summarization method based on RST. Here author assigned weights to the sentence in RStrees according to the utility, and cut out lower weight nodes. As a result the system generates complete, cohesive and readable summarization on the basis of relation between sentences in the original text.**

Graph Based Approach

Well known graph based algorithms are HITS and Google's PageRank [31]

- **HITS (Hyperlinked Induced Topic Search)**:
  It is a ranking algorithm for web page developed by Jon Kleinberg. It determines two set of scoresauthority: pages with large number of incoming links and hub: pages with large numbers of outgoing links [32].

$$HITS_H(V_i) = \sum_{v_j \epsilon Out(v_i)} HIT_A(V_j) HITS_H(V_i) = \sum_{v_j \epsilon Out(v_i)} HIT_H(V_j) \qquad (2.1)$$

  equqtion——————————————————————————-

- **Google's Pagerank Algorithm** :
  It is a ranking algorithm to determine quality of web pages. It is used by

Google to improve search result, named after Larry Page [33]. PageRank integrates both incoming and outgoing links into one single model, and therefore it produces only one set of scores:

$$PR(V_i) = (1 - d) + d * \sum_{v_j \epsilon In(v_i)} \frac{PR(v_j)}{|Out(v_j)|} \tag{2.2}$$

Where d is a parameter that is set between 0 and 1. Aardvark is a social search engine based on the village paradigm [34]. Miles Efron propose a page rank algorithm for Microblogs (e.g. Twitter) search [35].Daniel and Tunkelang proposed "a Twitter analog to PageRank[̈49]. It determines two set of scores Authority and Influence.

$$Influence(u) = \sum_{v \epsilon Follower(u)} \frac{1 + p * Influence(v)}{||Following(v)||} \tag{2.3}$$

Where Followers (.) is the set of people following a given user and Following is the set of people a given user follows and p is a realvalued number corresponding to the probability that a given tweet is retweeted.

Machine Learning Approach

Initially the system assumes that the features are independent. After that some feature dependent approaches are developed. The machine learning based summarization algorithms use techniques like NaveBayes, decision Trees, Hidden Markov Model.

- **NaiveBayes Methods :**

  NaiveBayes classifier, long a favorite punching bag of new classification techniques [36].A machine learning approach is based on three steps: Learning, Development and Test. Bayes rule takes feature of words and sentences as random events and relates to the conditional and marginal probabilities of those random events. According to Bayes rulen :

$$P(s \epsilon S | F_1, F_2, ..., F_k) = \frac{P(F_1, F_2, ..., F_k | s \epsilon S)}{P(F_1, F_2, ..., F_k)} \tag{2.4}$$

  Where s is a sentence from the document

S is the target summary

$(F_i)_{1\leq i\leq k}$ are features.

Andrew et al. compare Multivariate Bernoulli model and MultiNomial model and shown multivariate Bernoulli model performance is better. Rennie et.al discusses Multinomial Nave Bayes model and problems associated with it [37]. Mouratis et al. propose Discriminative Multinomial Bayesian Classifier, which increases the accuracy with a feature selection technique that evaluates the worth of an attribute by computing the value of the chisquared statistics with respect to the class [38].

- **Decision Trees : Decision tree is a classifier generated from training data to finding the feature in toptodown direction i.e. root to leaf node. Each node is generated based on the rules corresponding to the feature and this process is repeated until no further information gain is obtained.**

  Lin et al. assumed that the features are independent and applied decision tree algorithm for sentence extraction problem [39]. Data are used for this measurements provided by the TIPSTERSUMMAC. Collection of independent data is provided from SUMMARIST for assign score to sentences. SUMMARIST got same texts after applying each combination of functions, features and parameters. Some specific features are:

  Baseline: Scoring sentence by its position.

  Query signature : Normalized score of each sentence according to the number of uery words they contain.

  IR signature : most salient terms ranked by tfidf.

  Average lexical connectivity : Number of words shared with other sentences divided by the total number of sentences in the text.

  Numerical data: Boolean value 1 is given, if sentences contain numerical expression.

  Pronoun and adjective: Boolean value 1 is given if a sentence contain proper noun.

  Weekday and month : Boolean values 1 given to sentence if it contains weekdays and months.

Quotation : **Boolean value 1 given to sentences containing quote.**

When author applied these features to the query topic, they conclude that no single feature suffices for query based summaries.Kevin et al. proposed a model of sentence compression function using decision tree method [40].

- **Hidden Markov Model(HMM)**:

  It is a probabilisticfinitestate model for data. The structure of this model consists of number of states and transition between the states which is selected by the a priori of the domain.HMM is defined as follows [41]:

$$\lambda = (A, B, \pi) \tag{2.5}$$

S = Set of States, $S_1, S_2 ..... S_M$

V= Set of Output symbols, $V_1, V_2 ..... V_N$

Q= Fixed state sequence of length T,

O= Set of Observations of length T,

A= transition probability from State S_i **to** S_j**, denoted as** a_{ij} **, where**

$$a_{ij} = P(Q_T = S_j | Q_{T-1} = S_i) \tag{2.6}$$

B= Probability of Observation at k, produced from state S_j **denoted as,**$B = (b_i(k))$

$$b_i k = P(x_T = V_T | Q_T = S_i) \tag{2.7}$$

π**=Initial probability array, denoted as,**$\pi = [\pi_i]$

$$\pi_i = P(Q_1 = S_i) \tag{2.8}$$

There are two assumptions are made in the Markov model : Current state is dependent only on the previous state, and Output observation at time t is dependent only current state. It is also used for speech and handwriting recognition [42].

Zhou et al describe granularity refined DOM tree to extract detailed information combined with regular expression to extract fixed formative information [43]. They took training data set consists of address, room size, rent, area, telephone number, name etc, applied in DOM tree. Experiment showed better extraction results when it compared with RAPIER algorithm with same data sets.

- **MaximumEntropy Model** :

  A maximum entropy classifier can be used to extract sentences form documents. Osborne et al. specify that maximum entropy classifier showed better result in sentence extraction than naivebayes classifier when information is encoded in dependent features and independent features [44]. Maximum Entropy defined as [45]:

$$P(c|s) = \frac{1}{Z(s)} \exp(\sum_i \lambda_{i,c} f_{i,c}(s,c))) \tag{2.9}$$

  Where $z(s) = \sum_c \exp(\sum_i \lambda_i f_i(c,s))$**,is a normalized function,**
  $F_{i,c}$ **is a function for feature and c is class defined as:**

$$F_{i,c}(d,c') = \begin{cases} 1 & n_i(d) > 0 \textbf{ and } c' = c \\ 0 & \textbf{otherwise} \end{cases}$$

  The $\lambda_{i,c'}s$ **are feature weight parameters. The parameters values are used to maximize the entropy of the induced distribution based on the constraint.**
  Chieu et al. present a maximumentropy classification approach on a singleslot and multislot information extraction [46]. For singleslot task, they worked on seminar announcements. For this, they took several features such as,
  Unigram: The string of each word w is used as a feature. So is that of the previous word w-1 and the next word w+1.
  Bigram : The pair of word strings (w-2, w-1) of the previous two words is used as a feature. So is that of the next two words (w+1, w+2).
  Zone and InitCaps : Texts within the pair of tags ¡sentence¿ and¡/sentence¿ are taken to be one sentence. Words within sentence tags are taken to be in TXT zone. Words outside such tags are taken to be in a FRAG zone. This group of feature consists of 2 features (InitCaps, TXT) and (InitCaps,FRAG). For words starting with a capital letter (InitCaps), one of the 2 features (InitCaps,TXT) or (InitCaps,FRAG) will be set to 1, depending on the zone the word appears in.
  Zone and InitCaps of w-1 and w+1 : If the previous word has InitCaps, another feature (InitCaps, TXT)PREV or (InitCaps, FRAG)PREV will be set to 1.Same for the next word.
  Heading : Heading is defined to be the word before the last colon ¨: ¨ The system

will distinguish between words on the first line of the heading (e.g. Whofirstline) from words on other lines (Whootherlines). There is at most one feature set to 1 for this group.

First Word : This group contains only one feature FIRSTWORD, which is set to 1 if the word is the first word of a sentence.

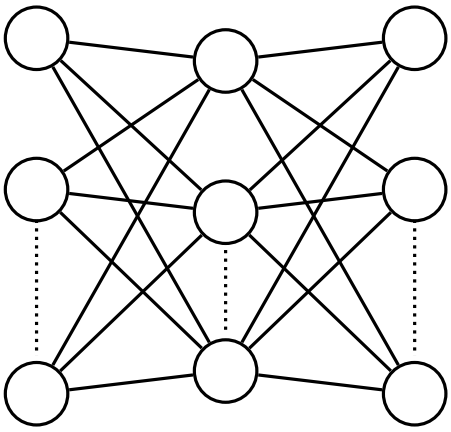
Time Expressions : If the word string of w matches the regular expression :[digit]+ :[digit]+, then this feature will be set to 1.

Names : If w has InitCapsand is found in the list of first names, the feature FIRSTNAME will be set to 1. If w-1 (or w+1) has InitCaps and is found in the list of first names then FIRSTNAMEPREV (FIRSTNAMENEXT) will be set to 1. Similarly for LASTNAME. For multislot task, they worked on Management Succession. The multislot IE system made iup of four components, such as, TextFiltering, Candidate Selection, Relation Classification, and Template Building. Author applied two benchmark data set for both task showed better accuracy in the Information extraction. Robert et al. compare number of algorithms for estimating the parameters of maximum entropy model including iterative scaling, gradient ascent, conjugate gradient, and variable metric methods [47]. Another new model; HiddenState Maximum Entropy (HSME) proposed which is based on fusion method for confidence measure [48].Concept of Maximum entropy model is also applied to Biological text terms boundary identification [1].

- **Neural Networks**:

  In 1997, Ruiz and Srinivasan [49] modeled a problem of recognizing MeSH term for a particular document . To solve this problem, they used backpropagation and counterpropagation networks.

  Backpropagation networks : Backpropagation network consists of two phases, one to propagate the input pattern and other to adapt the output by changing the weights in the network. The training procedure of a backpropagation network is iterative, with the weights adjusted after the presentation of each case. The



input (N_j) **and the output (O) of the network is defined as follows** :

$$N_j = \sum_i \omega_{ij} O_i + \Theta_j and O_j = \frac{1}{1 + e^{-N_j}} \tag{2.10}$$

Counterpropagation Network :

The Counterpropagation Network consists of an input layer, a hidden layer (also called Kohonen layer) , and an output layer (called Grossberg layer). The training process of consists of two steps, first, an unsupervised learning is performed by the hidden layer, then after the hidden layer is stable a supervised learning is performed by the outer layer. The formula of the hidden layer is :

$$\omega_{new} = \omega_{old} + \alpha(x - \omega_{old}) \tag{2.11}$$

Svore et al. approached a model based on neural nets, called NetSum for summarization and thirdparty datasets for features. Authors used as dataset of Wikipedia and CNN.com and applied a ranking algorithm, RankNet. The system performed well over the baseline of choosing the first n sentences of the document.

NTC (Neural Network Categorizer) [50] is a neural network model for representing documents into numerical vectors. It solved two problems, first : **it can classify documents with its huge dimensionality completely and second is,**

provides transparency about its classification. For text categorization, authors gather dataset from Newspage.com, 20NewsGroups, and Reuter . They applied four approaches, SVM, NB, KNN, BackPropagation, and compare with NTC for evaluation and got successful result as an approach to text categorization.

The Probabilistic Neural Network (PNN) was first proposed by Donald Specht in 1990. In 2009, Patrick et al. describe a modified model of PNN to solve the problem of Economic Activities Classification of Brazil [51].

Counterpropagation Network : Support vector machine is a learning method, developed by Vapnik et al. as stated in [52] Hirao et al. introduced a classification learning algorithm, Support Vector Machine (SVM) to categorize important or unimportant sentence in Single Document Summarization at Document Understanding Conference (DUC) [53]. Given training dataset (x_n, y_n)**, n=1 to n,**$x_j \epsilon R^n$ **and** $y_j \epsilon -1, +1$ **,where** x_j **is a feature vector of the jth sample and** y_j **is its class label (positive or negative). To rank sentences, they took features of sentences, such as Position of sentences, Length of sentences, weight of sentences, Similarity between Headline, Prepositions and verbs.**

They presented sentence ranking algorithm by SVM for multidocument summarization. To minimize redundancy, they applied Maximum Marginal Relevance (MMR) method. Novel features they used for ranking sentence are similar, the features they used for Single document summarization but in place of Similarity between Headlines, named entity is used.

In 2005, Minh et al. proposed a sentence extraction algorithm based on SVM ensemble classification to improve the accuracy for the data [54]. To correctly classify area in the training samples, they trained each SVM independently from the random chosen trained samples and to combine each machine, they used boosting strategy. To run this method, they implement Adaboosting algorithm to select training data for each individual SVM. Feature set were Location method, Length method, Relevant to title, term frequent and document frequent, cue phrase, distance of a word within a sentence.

Algebraic Approach

- **Latent Semantic Analysis: It is a algebraic statistical method to determine words and sentences which are semantically related. It creates a matrix representation by comparing semantic words .It is an algebraicbased Unsupervised approach. LSA produces measures of wordword, worddocument and documentdocument relations that are well correlated with several human cognitive phenomena involving association or semantic similarity [55].**
  Latent Semantic Indexing is a information retrieval method that project queries and documents into a space with "latent" semantic dimensions [56]. Singular Value Decomposition (SVD) is a method to find out the relations among very large number of words. It can reduce the noise and improve the accuracy.
  SVD : SVD of a matrix $A_{m\times n}$ defined as follows:

$$A = U_{m\times n} \times S_{r\times r} \times V^T_{r\times n} \tag{2.12}$$

Where U is Eigen vectors of AA^T,called term matrix,V is Eigen vectors of A^TA,called document matrix, and S is Eigen values of both A^TA and AA^T ,called diagonal matrix of nonzero singular values. Probabilistic Latent Semantic Analysis is a statistical model for word or document cooccurrences by the following scheme [57] :
Select a document d_i with probability $P(d_i)$, Pick a latent class z_k with probability $P(z_k|d_i)$, Generate a word ω_j with probability $P(\omega_j z_k)$. Where $P(d_i)$ is a probability that a word occur in a particular document.
$P(z_k|d_i)$ denote the probability distribution over a latent variable space,and $P(\omega_j z_k)$ denote the class conditional probability of a specific word conditioned on the unobserved class variable.
Meta Latent Semantic Analysis (MLSA) [58] improved accuracy model of LSA. It has the ability to create metaclusters by taking symbolic ontologies relevant for the analyzed collection of documents.Adaptive PLSA has the incremental learning capability to absorb the domain knowledge form new observed documents. It deals with domain mismatch for language processing

applications. To resolved updating problems, authors go through the foldingin, SVD recomputing, and SVD updating processes [59].

- **NonNegative Matrix factorization(NMF) : NMF is linear representation of nonnegative data applied to the set of multivariate ndimensional data vector. NMF model is defined as :**

$$A_{n\times m} \approx B_{n\times r} C_{r\times m} \tag{2.13}$$

Li et al. presented a multidimensional summarization framework based on sentence level semantic analysis (SLSS) and symmetric nonnegative matrix factorization (SNMF). SNMF can be 3factor nonnegative matrix factorization is defined as :

$$X \approx FSG \tag{2.14}$$

Where S provides lowrank matrix representation, F gives row clusters and G gives column clusters. After creating the clusters, authors rank the sentences based on sentence score. Sentence score can be measured as :

$$Score(S_i) = \lambda F_1(S_i) + (1-\lambda) F_2(S_i) \tag{2.15}$$

Where $F_1(S_i)$ measure the average similarity score between sentence S_i and all the other sentences in the cluster and $F_2(S_i)$ is the similarity between sentence and the given topic. λ is the weight parameter [60].

In 2009, Lee et al. [61] proposed an unsupervised NMF method to extract important sentences for automatic generic document summarization. Author claimed that NMF provide better performance in identifying subtopics of a document as compared with the methods using LSA because semantic feature vectors obtained using NMF have nonnegative values but in LSA method, it contain both positive and negative values.

- **Semi-Discrete Decomposition (SDD): SDD can be used in place of truncated SVD matrix is defined as [62]:**

$$A_k = \sum_{i=1}^{k} d_i x_i y_i^T \tag{2.16}$$

A rankk SDD requires the storage of k(m+n) values from the set-1,0,1 and k scalars.The scalar need to be only single precision because algorithm is self correcting.

To querybased text summarization, authors compared SVD and SDD based LSI methods on the MEDLINE dataset, requires only about half the query time, and requires less than onetwentieth the storage but to compute SDD approximation takes five times as long as computing the SVD approximation. Let $A \epsilon R^{m \times n}$ be a given matrix and let $w \epsilon R^{m \times n}$ be a nonnegative weighted matrix [62]. The weighted approximation problem is to find a matrix $A \epsilon R^{m \times n}$ that solves

$$min||A - B||_{\omega}^{2} \tag{2.17}$$

To overcome the problem of "curse of dimensionality", Vaclav et al. proposed a model Wordnet and Wordnet+LSI for dimension reduction [63]. Here SDD method is used to identify most conceptual terms. For identifying topic, SDD concept is used in two ways : to map the terms on synsets and use synset as input to the SDD for document and vectors.

2.1.3 Abstractive Text Summarization

In this method, machine need to understand the concept of all the input documents then produce summary with its own sentences. To accomplish this task, it go through these sub processes : information extraction, ontological information, information fusion and compression [64]. Machine uses linguistic methods to examine and interpret the text and then to find the new concepts and expressions to best describe it by generating a new shorter text that conveys the most important information form the original text document [65]. Witbrock et al. proposed a statistical approach model of nonextractive summarization process based on sentence compression. Main steps in this system are [66]:

a. Tokenization : Tokens may include not only the words, but additional information such as parts of speech tag, semantic tags applied to words, even phrases. Long

distance relationships between words or phrases in the document, positions of words or phrases, markup information obtained from the document such as existence of different font, etc. could be used. This preprocessing model is applied in both input documents and target documents.

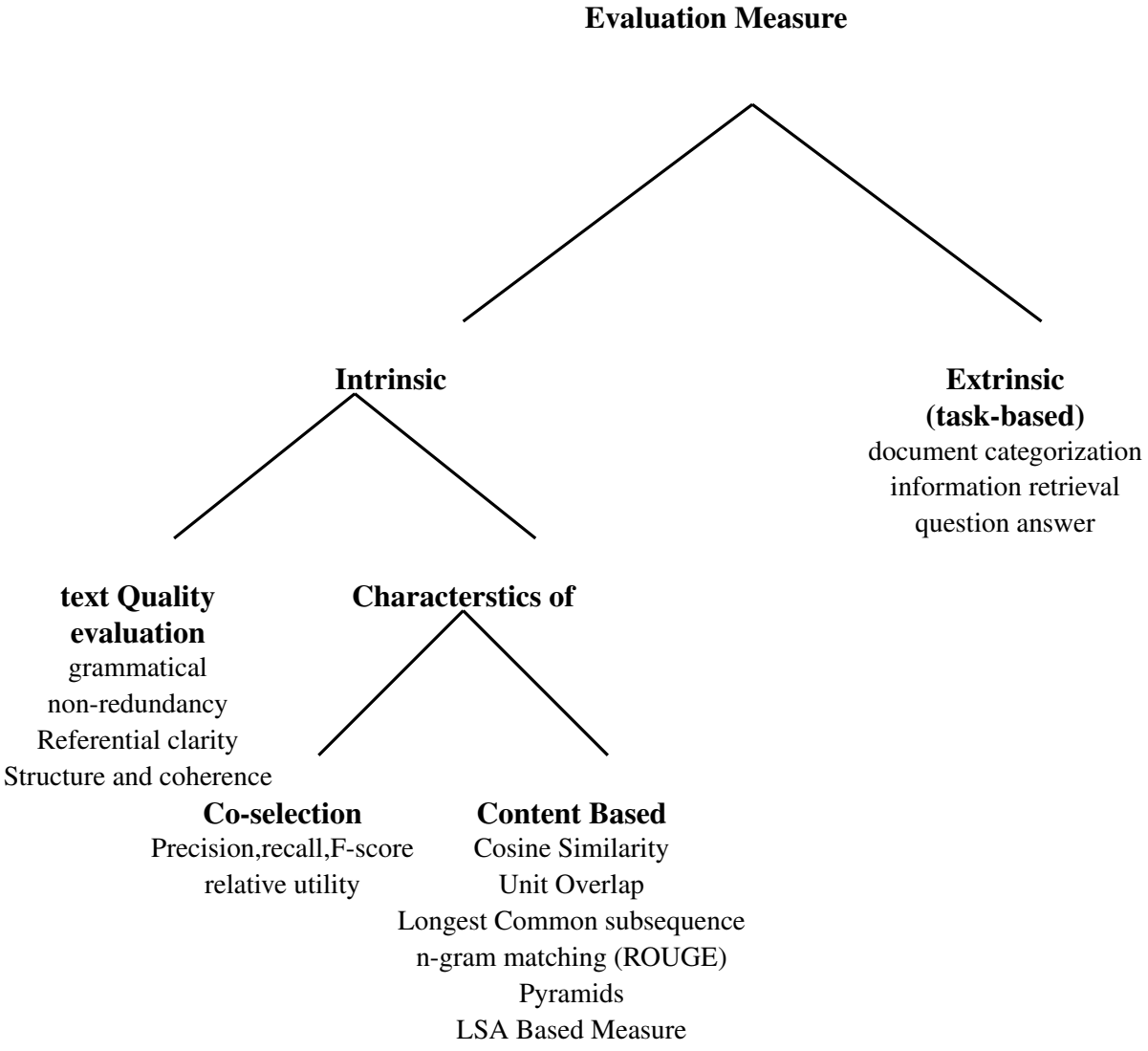
b. The statistical model is built describing the relationship between the source text units in a document and the target text units to be used in the summary of that document. It describes both the order and likelihood of appearance of the tokens in the target documents.

c. The statistical model generated information about user or task requirements, are used to produce the summary of a document.

2.2 Evaluation Measure

After creating automatic summary require to know, how useful it is. Whether it can fulfill the requirement for human or it is giving quality information or not. For this, automatic evaluation is done. TIPSTER Text Summarization Evaluation (SUMMAC), which was the first largescale, developerindependent evaluation of automatic text summarization system [67]. To evaluate a summary, baseline summaries need to create : single baseline summary for singledocument summarization and one baseline, lead baseline, coverage baseline summaries for multidocument summarization which is a difficult task [68]. Human evaluation task is expensive, very difficult and take more time. BLEU is a automatic evaluation of machine translation, inexpensive, quick and languageindependent, that correlates highly with human evaluation [69].There is no standard metric is defined for evaluation, which makes very hard to compare different systems and establish a baseline [70].

NIST did not define any official performance metric in DUC 2001 as stated by Lin (2002).Evaluation measures are categorized into two types, intrinsic and extrinsic evaluation. Intrinsic evaluation, judges the quality of the summary directly based on analysis in terms of some set of norms but extrinsic evaluation judges the quality of the summary based on the how it affects the completion of some other task. The taxonomy of evaluation measure as stated in [71] shown in figure ??.


Evaluation Measure
Intrinsic
Extrinsic
(task-based)
document categorization
information retrieval
question answer
text Quality
evaluation
grammatical
non-redundancy
Referential clarity
Structure and coherence
Characterstics of
Co-selection
Precision,recall,F-score
relative utility
Content Based
Cosine Similarity
Unit Overlap
Longest Common subsequence
n-gram matching (ROUGE)
Pyramids
LSA Based Measure

- **Text Quality Measure** : **Grammaticality** : **The summary should not contain any grammatical error like punctuation errors or incorrect words. Nonredundancy** : **the summary should not contain any redundant information. Referencequality** :**The reference in the summary clearly matched with the known object. Coherence and structure** : **The summary should have good structure and the sentences are coherently related.**

- **Coselection Measures**: **Here sentences are extracted from the created summary and evaluate against the human selection. The metrics of coselection are Precision, Recall and Fscore. Precision** : **Precision defined as the proportion of retrieved documents that are relevant [72] or Common extracted the number sentences from system and human choice summary divided by number of sentences extracted from system summary.**

$$Precision = \frac{SystemSentences \cup HumanJudgesChoiceSentences}{SystemSentences} \tag{2.18}$$

Recall : **Recall is defined as the proportion of relevant documents that are retrieved [89] or common number sentences extracted from system summary and human choice summary divided by number of sentences extracted from human choice summary.**

$$Precision = \frac{SystemSentences \cup HumanJudgesChoiceSentences}{HumanJudgesChoiceSentences} \tag{2.19}$$

F Score : **FScore is a statistical method that combines precision and recall. F-score is defined as harmonic average of precision and recall. Its value lies between 0 and 1 where 1 is best value.**

$$FScore = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{2.20}$$

Another formula for FScore for measuring the FScore :

$$FScore = \frac{(\beta^2 + 1) \times Precision \times Recall}{\beta^2 \times Precision + Recall} \tag{2.21}$$

Where β is a weight value not equal to zero. For $\beta > 1$, it indicate Precision is more important and for $\beta < 1$ indicate Recall is more important.

Relative Utility :Relative Utility measure to overcome the problem of the Precision and recall based evaluation as stated in [73]. Suppose a manual summary contain sentences 1, 2, 3, and 4 from a document. There are two systems S1 and S2, creates summaries consisting of sentences 1, 2, 4 and 1, 2, 3.It can be possible that two sentences in one document are equally important. Using Precision and Recall, S_1 can rank higher than S_2. Judges to judges, ranking of sentences are varies. If a particular sentence ranked 8 by judge 1 and same sentence is ranked 10 by judge2, then utility score of that sentence is 0.8 ($\frac{8}{10}$). To calculate Relative Utility, a number of judges ($N \geq 1$) are asked to assign utility score to all sentences in the document. The top e number of sentences is extracted according to utility score. Relative Utility of a system is calculated as :

$$RelativeUtility = \frac{\sum_{i=1}^{n} \delta_j \sum_{j=1}^{N} \lambda_{ij}}{\sum_{i=1}^{n} \eta_j \sum_{j=1}^{N} \lambda_{ij}} \tag{2.22}$$

Coselection based evaluation focused on summaries where sentences are extracted.

- **Content based Measure : Content based evaluation mainly focuses on extracted summaries where comparison is done among words. Measures of Contentbased evaluation are : Cosine similarity, unit overlap, longest common subsequence, ROUGE score, and pyramid. Cosine Similarity :**

$$sim(D_1, D_2) = \frac{\sum_i d_{1i} d_{2i}}{\sqrt{\sum_i (x_i)^2} \sqrt{\sum_i (y_i^2)}} \tag{2.23}$$

Where D_1 and D_2 are two documents represented using a vector space model and d_i is a term weight for $word_i$.

Unit Overlap : Unit Overlap is defined as :

$$overlap(X, Y) = \frac{||X \cap Y||}{||X|| + ||Y|| - ||X \cap ||} \tag{2.24}$$

Where X and Y are text representations based on sets. Here $||S||$ is the size of set S. Longest Common Subsequence : LCS finds longest common subsequence of X and Y. It can be calculated as [74] :

$$2 \times lcs(X, Y) = length(X) + length(Y) - edit_{di}(X, Y) \tag{2.25}$$

Where $length(X)$ and $length(X)$ are length of the string X and Y respectively and $edit_{di}(X, Y)$ is the edit distance between X and Y.

ROUGEN : Ngram CoOccurrence Statistics: ROUGEN is an ngram recall between a candidate summary and a set of reference summaries. ROUGE-N is computed as follows :

$$ROUGE - N = \frac{\sum_{s \epsilon ReferencesSummaries} \sum_{gram_n} Count_{match}(gram_n)}{\sum_{s \epsilon ReferencesSummaries} \sum_{gram_n} Count(gram_{ns})} \tag{2.26}$$

Where $gram_n$ and $Count_{match}(gram_n)$ is the maximum number of ngrams cooccurring in a candidate summary and a set of reference summaries, and n stands for length of the ngram.

In case of multiple references, pairwise summarylevel ROUGEN between a candidate summary s and every reference,r_i, in the reference set. ROUGEN can be computed for multiple reference as follows :

$$ROUGE - N_{multi} = argmax_i ROUGE - N(r_i, s) \tag{2.27}$$

ROUGE can be computed based on longest common subsequence, known as ROUGEL. Fmeasure base on LCS can be computed as :

$$F_{lcs} = \frac{(1 + \beta_2 R_{lcs} P_{lcs}}{R_{lcs} + \beta^2 P_{lcs}} \tag{2.28}$$

where $R_{lcs} = \frac{LCS(X,Y)}{m}$ and $P_{lcs} = \frac{LCS(X,Y)}{n}$, X is a reference summary sentence of length of m and Y is a reference summary sentence of length n. ROUGEL does not require consecutive matches but insequence matches that reflect sentence level word order as ngram and it automatically includes longest insequence

common ngrams; therefore no predefined ngram length is necessary. Sentence level LCS based can be applies to the summarylevel. In this process, it take union LCS matches between reference summary sentence,r_i,and every candidate summary sentence,c_j . Fmeasure can be computed as [75] :

$$F_{lcs} = \frac{(1+\beta^2)R_{lcs}P_{lcs}}{P_{lcs}+\beta^2 P_{lcs}} \tag{2.29}$$

where $R_{lcs} = \frac{\sum_{i=1}^{u} LCS_{\cup}(r_i,c)}{m}$ **and** $P_{lcs} = \frac{\sum_{i=1}^{u} LCS_{\cup}(r_i,c)}{n}$**,and number of sentences containing a total number on m words in reference summary and v number of sentences containing a total number of n words in candidate summary.**

Pyramids : To identify relevant information from s document or set of documents. It is based on Summary Content Unit (SCU). A SCU is a semantically atomic unit representing a single fact, but is not tied its lexical realization [76]. Let be the number of SCUs in the summary that appear in tier Ti, and X is the total number of SCUs in the summary. Total SCU weight can be computed as :

$$D = \sum_{i=1}^{n} i \times D_i \tag{2.30}$$

This SCU weight is then normalized by the optimal content score for a summary X SCUs.The optimal content score is computed as :

$$Max = \sum_{i=j+1}^{n} i|T_i| + j(X - \sum_{i=j+1}^{n} n|T_i|) \tag{2.31}$$

where $j = max(\sum_{i=1}^{n} |T_i| \geq X)$ **This pyramid score lies between 0 and due to normalization.**

LSA Based measure : It has the ability to capture the most important topics is used by the two evaluation metrics proposed by Steinberg et al. It evaluates a summary quality via content similarity between a reference document and the summary. The quality is measured by the similarity between the matrix U derived from the SVD performed on the reference document and the matrix U derived from the SVD performed on the summary. There are two similarity measures proposed : Main Topic Similarity and Term Significance Similarity.

- **Task-Based Measure** : **Task based evaluation focus on the quality of a summary according to the fulfillment of a user. It requires more effort than intrinsic evaluation. Approaches of taskbased summarization evaluation are** : **Document categorization, information retrieval and question answering.**

  Document categorization : **It determines whether the summary is effective in capturing whatever information in the document is needed to correctly categorize the document. Categorization can be done by human judges or automatic classifier. By comparing the upper and lower bounds of the error generated by a classifier and one that by a summarizer, we can compare the system performance. The evaluation metrics of categorization are** : **Precision and recall. Precision in this context is the number of correct topics assigned to a document divided by the total number of topics assigned to the document. Recall is the number of correct topics assigned to a document divided by the total number of topics that should be assigned to the document.**

  Information Retrieval : **It is a appropriate taskbased evaluation of a summary quality. Relevance Correlation is an IR based measure for assessing the relative decrease in retrieval performance when moving from full documents to summaries. It measures the quality of summaries by comparing how well the summary and full document does.There are several methods for measuring the similarity of rankings. One such method is Kendall's tau and another is Spearman's rank correlation. Relevance correlation r is defined as the linear correlation of the relevance scores (x and y) assigned by two different IR algorithms on the same set of documents or by the same IR algorithm on different data sets** :

$$r = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2}\sqrt{\sum_i (x_i - \bar{x})^2}} \tag{2.32}$$

  Here $\bar{x}$ and $\bar{y}$ are the means of the relevance scores x and y for the document sequence respectively.

  Question Answer : **Here Authors take a test which consists of multiple choices, with a single answer to be selected from answer shown alongside each question. Authors measured how any of the questions the subjects answered correctly**

under different conditions by compare with professional answer.

2.3 Keyword Extraction

To extract important information or sentences, high quality keyword plays crucial role as per user requirement. They help users to search information more efficiently. Keyword extraction can be used in many applications, such as text summarization, clustering, classification, topic detection, etc [77]. Due to growth of online information it is difficult for human beings to accomplish their task in the field of natural language processing in stipulated time. Extracting high quality keywords automatically are expensive and time consuming. This shows keyword extraction is challenging problem in the area of natural language processing especially in the context of global languages in acceptable time.

Frank et al. investigate keyword extraction algorithm as a supervised learning algorithm [78]. They also introduced KEA algorithm for keyword extraction.tf-idf method is used for feature calculation [79] and it performed well. In 2000, Turneyet al. used decision algorithm and genetic algorithm for keyword extraction [80].Kerner *et al.* [81]investigate tf-idf is very effective in extracting keywords for scientific journals. Keyword extraction also solved as unsupervised approach task shown by Lie et al [82].Barker *et al.* discusses a key phrase extraction system that scores to noun phrases based on frequency and length and it also filter some noise from the set of top scoring keyphrases [83].Daille et al. applied linguistic knowledge to identify noun phrases for both in English and French terms [84]. They used statistical methods to score good terms.

Keyword extraction methods can be divided into different categories based on approaches:-

Statistical approach : These methods are simple and do not need the training data. The statistics information of the words can be used to identify the keywords in the document. It includes n-gram, word frequency, tf-idf, and word co-occurrence methods.Burnett *et al.* used n-gram to identify index terms in document [85]. Cohen investigates n-gram count method to extracting highlights from the document [86].In 1957,Luhn described statistical approach that a sentence gives useful measurement

of significance, if frequency of particular term (or word) is high in an article. Frequencies of pair of words is high in the documents then term co-occurrence value is high [87].

Linguistic approach: These approaches use the linguistics feature of the words mainly, sentences and document. The linguistics approach includes the lexical analysis, syntactic analysis etc.Lexical chain is a method of identifying set of words which are semantically related. WordNet is used for measuring of conceptually similarity and relatedness information from document [88], [27]. Hulth used syntactic features for extracting keywords. To give an idea about pattern, frequently occurring keywords present in the training data are adjective noun (singular or mass), noun noun (both singular or mass), adjective noun (plural), noun (singular or mass) noun (plural) and noun (singular or mass) [89].Other researchers used lexical cohesion method for keyword extraction such as, Brazilay et al, Angheluta et al. [90], [91].

Machine learning approach: It includes methods like naive bayes, support vector machine, etc. Bayesian decision theory based on tradeoffs between the classifications decisions using probability and the costs that accompany those decisions [92].It examined that it is less favourable due to large training data set.Zhang et al.defined three categories of keywords, such as 'good keyword', 'indifferent keyword' and 'bad keyword'. They applied support vector machine as a classification model for keywords [93].

Other approach: It includes method that uses some heuristic knowledge, such as the position, length, html tag etc. Position of the word appears defined by its position normalized by the total number of words in the document. Keywords are extracted based on the maximum length and highest salience score of the sentences [94]. Humphreys investigate on HTML keyword extractor. It is based on phrase rate that includes word rate, docrate, ratephrases and selector [95].It is especially suitable for online keyword aid.

Chapter 3

Comparison between Performances of two Keyword Extraction Methods

Here I describe our work on comparison between performance of keyword extraction methods that are most popular TF-IDF method and another is based on Helmholtz Principle. Here I propose a algorithm based on Helmholtz Principle to get meaningful words in stipulated time.

3.1 Term Frequency-Inverse Document Frequency (TF-IDF):

tf-idf (term frequency-inverse document frequency) weight identify importance of words to a document collection. Important keywords that appear frequently in a document, but that don't appear frequently in the remainder of the corpus [23].The tf measures the number of times a word appears in the current document which can reflects the frequency of the word in this article, while the idf reflects the number of documents in which the word occurs. When the word is more frequent in the sentence but less frequent in the whole document, the tf-idf value is higher.tf-idf is defined as:

$$tf - idf = tf \times idf \tag{3.1}$$

$$idf(i) = \log \frac{n}{n(i)} \tag{3.2}$$

Where tf= number of times term i occur in document,
n= number of documents in the corpus and
n(i)= number of documents in which the word i occurs
tf-idf assigns to term t a weight in document d that is

- **highest when t occurs many times a small within a small number of documents;**
- **lower when the term occurs fewer times in a document, or occurs in many documents;**
- **lowest when the term occurs in virtually all documents**

3.2 Optimization of Meaningful Keywords Extraction using Helmholtz Principle

Jon Kleinberg present a formal approach for modeling "bursts," so that they can be robustly and efficiently identified [24]. According to a basic principle of perception due to Helmholtz, an observed geometric structure is perceptually meaningful if it has a very low probability to appear in noise. As a common sense statement, this means that events that could not happen by chance are immediately perceived. For example, a group of five aligned dots exists in both images in Figure ??, but it can hardly be seen on the left-hand side image. Indeed, such a configuration is not exceptional in view of the total number of dots. In the right-hand image we immediately perceive the alignment as a large deviation from randomness that would be unlikely to happen by chance.

In the case of textual, sequential or unstructured data, Balinsky *et al.* derive qualitative measure for such deviations. Suppose we are given a set of N documents $D_1, D_2, ..., D_N$(containers) of the same length [26]. Let W be some words inside these N documents. Assume that the word W appears K times in all N documents and let us collect all of them into one set $S_w = w_1, w_2, ..., w_N$.Let us denote by C_m,a random

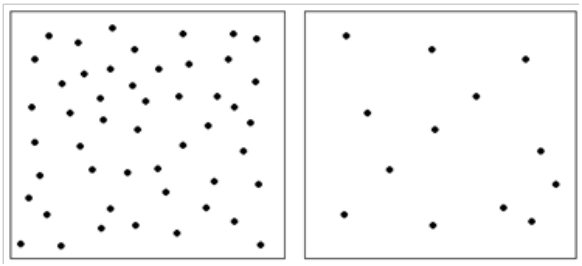


Figure 3.1:

variable that counts how many times an m-tuple of the elements of S_w appears in the same document. Now we would like to calculate the expected value of the random variable C_m under an assumption that elements from S_w are randomly placed into N containers. Form different indexes $i_1, i_2, ...i_m$ between 1 and K i.e. $1 < i_1, i_2, ..., i_m < k$ a random variable

$$X_{i1,i2,...,im} = \begin{cases} 1 & \textbf{if } w_{i1,...,w_im} \textbf{ are in same document} \\ 0 & \textbf{otherwise} \end{cases}$$

The function C_m,

$$C_m = \sum_{1 \le i_1 < i_2 < i_m \le} X_{i1,i2,...,im} \tag{3.3}$$

and that expected Value $E(C_m)$ is sum of expected values of all $X_{i1,i2,...,im<k}$,:

$$E(C_m) = \sum_{1 \le i_1 < i_2 < i_m \le} E(X_{i1,i2,...,im}) \tag{3.4}$$

Since $X_{i1,i2,...,im}$ has only values zero and one, the expected value $E(X_{i1,i2,...,im})$ is equal to the probalbility that all $w_{i_1}, w_{i_2}....w_{i_m}$ belong to the same document, i.e.

$$E(X_{i1,i2,...,im}) = \frac{1}{N^{m-1}} \tag{3.5}$$

From the above identities, we can see that

$$E(C_m) = \frac{K!}{m!(K-m)!} \cdot \frac{1}{N^{m-1}} \tag{3.6}$$

We define $\frac{K!}{m!(K-m)!} \cdot \frac{1}{N^{m-1}}$ as the number of false alarms(NFA) of a m-tuple of the word W.

The word W appears m times in the same document, then we define this word as a meaningful word if and only if its NFA is smaller than 1.

If NFA is less than ϵ, we say that W is ϵ meaningful.

Algorithm 3.1 Calculate NFA(N,L,M,k,m)

Input: Store each document into an array from D_1 to D_N

```
Set corpus=[];
Add all the documents D1 to D2 into corpus Array;
L ← length(corpus);
Set W=[];
for i := 0 to L do
    W = append(Uniquewords(corpus));
end for
K=[];
for i := 1 to lenth(W) do
    Set counter=0;
    for j := 1 to L do
        if W[i] == corpus[j] then
            counter=counter+1;
        end if
    end for
    K[j]=append(counter);
end for
B;
Window Size
x=[],y=[],z=[];
for i := 1 to N do
    l ← Di
    for j := 1 to B do
        X[j]=appendDi[j];
        if B ≤ l then
            for k = (j + 1) to (B + 1) do
                y[k] = append(Di[k]);
                x=GetIntersection(x,y);
                j=j+1;
                B=B+1;
            end for
        end if
    end for
end for
M ← L/B;
for D(i = 1)toD(i = N) do
    m=[];
    for j = 1 to length(x) do
        counter =0;
        for k = 1 to l do
            if x[i] == Di[k] then
                counter=counter+1;
            end if
        end for
        m[j] = append(counter);
    end for
end for
```

Set Word=[];
for i = 1 to lengh(x) **do**
  for j = 1 to lengh(W) **do**
    if $x[i] == W[j]$ **then**
      p=K[j];
      q=m[i];
      if $\frac{p!}{q!(p-q)!} \times \frac{1}{M^(q-1)}$ ¡1 **then**
        $Word = append(x[i])$
      end if
    end if
  end for
end for

Figure 3.2:

In a case of one document or data stream it can be divided into a sequence of disjoint and equal size blocks and performs analysis for the documents of equal size. Since such a subdivision can cut topics and is not shift invariant, the better way is to work with a "moving window". An example of moving window is shown in Figure 3.2.

More precisely, if we are given a document D of the size L and B is a block size. We define N as$[\frac{L}{B}]$. For any word W from D and any windows of consecutive B words let m count number of Win this windows and K count number of W in D. If $NFA < 1$, where

$$\frac{K!}{m!(K-m)!} \cdot \frac{1}{N^{m-1}} < 1 \tag{3.7}$$

then add W to a set of keywords and say that W is meaningful in these windows. In the case of one big document that has been subdivided into subdocuments or sections, the size of such parts are natural selection for the size of windows.

In real life examples it cannot be possible that a corpus of N documents $D_1, D_2, ..D_N$ have the same length. Let l_i denotethe length of the documentD_i. We followed some strategies for creating a set of keywords, such as:

- **Subdivide the set $D_1, D_2, ..D_N$into several subsets of approximately equal size documents, and perform analysis above for each subset separately.**
- **'Scale' each document to common length l of the smallest document. More precisely, for any word we calculate as $K = \sum_{i=1}^{N} [\frac{m_i}{l}]$, where [x] denotes an integer part of a number x and m_i counts the number of appearances of the word W in a document D_i.For each documentD_i, we calculate the NFA with this K and the new $m_i \leftarrow [\frac{m_i}{l}]$.All words with NFA¡1 comprise a set of keywords.**

Chapter 4

Evaluation and Results

After brief description of the Text summarization systems, in this chapter I have collected information on automatic text summarization systems. Also I summarize experimental evaluation of two keyword extraction methods presented in the previous chapter. First, in Section 4.1, I have given short description of more than 50 automatic text summarization systems. In Section 4.2, I show the experimental result of comparison between two keyword extraction methods for automatically extracting meaningful keywords and their execution time. Also I present a model how proposed algorithm is implemented.

4.1 Text Summarization Systems

In order to understand what each column means, the following information is provided in Table 4.1. In first column (SYS, [REF], YEAR) the name of the system with its reference and year is written, the second column (INPUTs) distinguish between single document or multi-document summarization (both inputs can be possible). Third column (DOMAIN) indicates genre of the input that is, whether it is designed for specific domain or for non-restricted domain. Next column (FEATURES) describes the characteristics and techniques used in each system. Fifth column (EVALUATION) represents what the authors evaluate to get required output. Next column (METRICS) represents the metrics used in each system. The last column (OUTPUT) represents whether the summary generated is either an extract or an

abstract.

Table 4.1: Overview of Text Summarization Systems

SYS, [REF], YEAR	INPUTs	DOMAIN	FEATURE	EVALUATION	METRICS	OUTPUT
ADAM, [96], 1975	Single Document	Domain Specific : Chemistry	Semantic codes and sentence rejection or selection. Syntactic codes and coherence.	Program speed and Abstract size	Human Judgements	Indicative Abstracts
ANES, [97], 1995	Multi document	Domain specific: news	Statistical corpus analysis, signature word selection, sentence weighting, and sentence selection	acceptability of the summaries Summary rejection analysis Retrieval effectiveness evaluation.	Recall and Precision	Indicative Abstracts

Continued on next page ...

Table 4.1 – continued from previous page . . .

SYS, [REF], YEAR	INPUTs	DOMAIN	FEATURE	EVALUATION	METRICS	OUTPUT
Dimsum, [98], 1997	Single document	Domain Independent system	DUses NLP tool to extract multi-word phrases automatically Acquisition of some domain knowledge from a large corpus by calculating idf values for selecting signature words, deriving collocations statistically, and creating a word association index to capture lexical cohesion of signature words through name aliasing with the NameTag tool, synonyms with WordNet, and morphological variants with morphological pre-processing. Experimented with two stage combining summarization features: Batch Feature Combiner and Trainable Feature Combiner	generic summary could substitute for a full-text document	Precision and Recall	Extracts

Continued on next page . . .

Table 4.1 – continued from previous page . . .

SYS, [REF], YEAR	INPUTs	DOMAIN	FEATURE	EVALUATION	METRICS	OUTPUT
SUMMONs, [99], 1998	Multi document	Domain specific: Online news	It extracts data from the different sources and then combines it into a conceptual representation of the summary. Conceptual representation of the summary is passed to the lexical chooser then it passed through a sentence generator using FUF/SURGE language generation system.	To determine quality of generated summary under taskbased evaluation	Coverage	Extracts Abstracts
SUMMARIST, [100], 1998	Multi document	Domain specific: news	It combines robust NLP processing with symbolic world knowledge. It performs Topic identification, Topic Interpretation and Summary generation	Quality of summary.	Compression Ratio, Retention Ratio, Precision and Recall	Extracts

Continued on next page . . .

Table 4.1 – continued from previous page . . .

SYS, [REF], YEAR	INPUTs	DOMAIN	FEATURE	EVALUATION	METRICS	OUTPUT
Marcu, [101], 1999	Single document	Domain specific: news	Discourse-based Summarizer. it uses the rhetorical parsing algorithm to determine discourse structure of the text of given input, Determine partial ordering on the elementary and parenthetical units of the text.	To determine adequacy for summarizing texts for discourse-based methods	Precision and Recall	Extracts

Continued on next page . . .

Table 4.1 – continued from previous page . . .

SYS, [REF], YEAR	INPUTs	DOMAIN	FEATURE	EVALUATION	METRICS	OUTPUT
MultiGen, [102], 1999	Multi document	Domain specific: news	The content planner finds an intersection of phrases by comparing the predicate argument structures, -also orders selected phrases and arguments with the information needed for clarification -produce fluent sentences that combine these phrases, arranges them in novel contexts. -to avoid redundant information in the summary, it intersects the theme sentences to identify the common phrases to generate new sentence. -used FUF/SURFE language generator.	To identify common phrases throughout multiple sentences for content selection stage.	Precision and Recall	Abstracts

Continued on next page . . .

Table 4.1 – continued from previous page . . .

SYS, [REF], YEAR	INPUTs	DOMAIN	FEATURE	EVALUATION	METRICS	OUTPUT
Chin & Len, [103], 2000	Multi document	Domain specific: news	It proposed a multilingual summarizer. -used English and Chinese language generator. -contains monolingual and multilingual clustering. It finds matching among the clusters in different languages in multilingual cluster. -three kinds of linguistic knowledge- punctuation marks, linking elements and topic chains. -find the similarity among meaningful units in the articles.	Perform similarity of Meaningful Units.	Precision Rate, Recall Rate	Extracts

Continued on next page . . .

Table 4.1 – continued from previous page ...

SYS, [REF], YEAR	INPUTs	DOMAIN	FEATURE	EVALUATION	METRICS	OUTPUT
MEAD, [104], 2001	Multi document	Domain specific: news articles	MEAD used three features such as centroid score, position, and overlap with first sentence. Uses LT-POS software to mark sentence boundaries automatically It discards most similar sentences and also considers length of the sentence. Experimented with Cross-Document Structure Theory (CST)	Relationship between pair articles	Human Judgments	Extracts

Continued on next page ...

Table 4.1 – continued from previous page ...

SYS, [REF], YEAR	INPUTs	DOMAIN	FEATURE	EVALUATION	METRICS	OUTPUT
NewsInEssence, [105], 2001	Multi document	Domain specific: news articles	It summarize topic-based cluster of articles. I find articles by traversing links from the page and add into the cluster of similar articles by going to the search engines. NewsTroll determines interesting URL to fetch new page. CST is used to find relations between clusters	blank	Unknown	Extracts

Continued on next page ...

Table 4.1 – continued from previous page . . .

SYS, [REF], YEAR	INPUTs	DOMAIN	FEATURE	EVALUATION	METRICS	OUTPUT
WebInEssence, [106], 2001	Multi document	Domain-independent system	Uses a Web-based multi-document summarization and recommendation system. - centroid-based technique is used. Four major modes of operations are: Generic search, Generic search+Summarization, Generic search+Clustering+Summarization, Personalize mode. It uses a personalized search engine called MySearch.	To identify most relevant clusters. To improve scalability, readability and Usability	Cluster Score, Catching document, Catching query results, Keyword in the context etc	Extracts and Personalized summaries

Continued on next page . . .

Table 4.1 – continued from previous page ...

SYS, [REF], YEAR	INPUTs	DOMAIN	FEATURE	EVALUATION	METRICS	OUTPUT
NeATS, [107], 2002	Multi document	Domain specific: news articles	Techniques used: sentence position, term frequency, topic signature term clustering. To improve cohesion and coherence, stigma word filters and time stamps are used. Webclopedia's ranking algorithm is used to rank sentences. To records the relative importance of sentence positions.	In DUC-01.Evaluate most relevant sentences of the system summary and compare with human judgment.	Precision, Recall, F-Measure.	Extracts

Continued on next page ...

Table 4.1 – continued from previous page ...

SYS, [REF], YEAR	INPUTs	DOMAIN	FEATURE	EVALUATION	METRICS	OUTPUT
Columbia, [108], 2002	Multi document	Domain specific: news articles	It is a composite of two systems, MultiGen and DEM for generating single document summaries and multi-document summaries respectively. Statistical parameters are used to extract sentences. Ability to generate extractive and abstractive summaries	Quality of summary compared with human judgment.	Precision, Recall	Extracts & Abstracts

Continued on next page ...

Table 4.1 – continued from previous page . . .

SYS, [REF], YEAR	INPUTs	DOMAIN	FEATURE	EVALUATION	METRICS	OUTPUT
GLEANS, [109], 2002	Multi document	Unknown	It maps all the documents into database-like representation. -classifies into four categories: single person, single event, multiple event, and natural disaster. -it generates a short headline using a set of predefined templates. -generate summaries by extracting sentences from the database.	Evaluate on DUC-2002 corpus. Determine error of different categories	Coverage score, grammatical effect, Coherence, Cohesion	Headlines, Extracts & Abstracts

Continued on next page . . .

Table 4.1 – continued from previous page . . .

SYS, [REF], YEAR	INPUTs	DOMAIN	FEATURE	EVALUATION	METRICS	OUTPUT
GISTexter, [110], 2002	Single and Multi document	Domain specific: news articles	For single-document summarization, sentence extraction is done and filters out unnecessary information. For multi-document summarization, when Topic is known: CICERO Information Extraction identifies all the necessary information used in the multi-document summary. Topic is not known: modeling the topic in an ad-hoc manner to generate the summary.	Evaluated on DUC-2002 corpus and measure the overlap between systems generated summaries and the gold standard summary, human made summary	Precision and Recall.	Headlines, Extracts & Abstracts

Continued on next page . . .

Table 4.1 – continued from previous page . . .

SYS, [REF], YEAR	INPUTs	DOMAIN	FEATURE	EVALUATION	METRICS	OUTPUT
NTT, [53], 2002	Single document	Unknown	Features for sentence extraction are: sentence position, length , weight, similarity between headlines prepositions, verbs. To classify sentence, Support Vector Machine and Machine Learning Algorithm is used.	Experimented with DUC-2002 data to evaluate quality of the generated summary	Mean Coverage, Length -Adjusted Coverage, Readability metrics	Extracts
SumUM, [111], 2002	Multi document	Domain-specific technical articles	Explore the issues of dynamic summarization. -it composite of shallow syntactic and semantic analysis, concept identification, and text regeneration processes	made in intrinsic or extrinsic fashions. Intrinsic evaluation measures the quality of the summary and Extrinsic how helpful a summary is.	Precision and F-Score.	Abstracts

Continued on next page . . .

Table 4.1 – continued from previous page ...

SYS, [REF], YEAR	INPUTs	DOMAIN	FEATURE	EVALUATION	METRICS	OUTPUT
Newsblaster, [108], 2002	Multi document	Domain specific: news articles	It is an on-line news summarization system.0 Articles are clustered using Topic Detection and Tracking (TDT) system. It uses agglomerative clustering method and log-linear statistical model to group similar features. Features are: terms, noun phrase heads and proper nouns. Thumbnails of images are displayed.	Used DUC corpus. It automatically summarizes the single-event and multi-event documents. DEMS (Dissimilarity Engine for Multidocument Summarization) system is used for biographical documents.	Precision and Recall.	Extracts

Continued on next page ...

Table 4.1 – continued from previous page . . .

SYS, [REF], YEAR	INPUTs	DOMAIN	FEATURE	EVALUATION	METRICS	OUTPUT
Robust Generic and Query based Summarization, 2003	Single document [112],	Unknown	GATE components produced by ANNIE, combined with well established statistical techniques. Supports generic and query based summarization.	Determine the quality of the output document compared with human made documents. Identified best feature combination	Precision, Recall and F-Score	Extracts
Lethbridge, 2003	[113], Single and Multi document	Unknown	It can produce very short summaries. -used TDT and TREC clusters techniques. Lexical information is used TDT topic detection. Relevant sentences were extracted using scoring process.	Used DUC 2003 documents. It evaluates the quality of output documents.	mean coverage, median coverage, mean quality questions, mean length-adjusted coverage.	Extracts

Continued on next page . . .

Table 4.1 – continued from previous page . . .

SYS, [REF], YEAR	INPUTs	DOMAIN	FEATURE	EVALUATION	METRICS	OUTPUT
Copeck et al.,[114], 2003	Single document	Unknown	To detect sentences and paragraphs, it remove tables, abstracts, reference list , and page headers and footers. -First extracts key phrases from the document, than depending on most pertinent key phrases, it picks sentences. LT-POS chunking parser is used to generate headlines. TDT technique is used for topic specification.	DUC 2003 data set is used. Length of summaries, proper-quality substring database, duplicates matches of viewpoint phrase.	mean coverage, median coverage.	Extracts

Continued on next page . . .

Table 4.1 – continued from previous page ...

SYS, [REF], YEAR	INPUTs	DOMAIN	FEATURE	EVALUATION	METRICS	OUTPUT
Erkan et al.,[115], 2004	Multi document	Unknown	It is an extractive summarization environment. It consists of three steps: feature extractor, feature vector, and reranker. Features are, Centroid, Position, LengthCutoff, SimWithFirst, LexPageRank, and QueryPhraseMatch.	Evaluate in DUC 2004. It evaluates overall system performance.	ROUGE-1, ROUGE-W	Extracts

Continued on next page ...

Table 4.1 – continued from previous page . . .

SYS, [REF], YEAR	INPUTs	DOMAIN	FEATURE	EVALUATION	METRICS	OUTPUT
UAM,[116], 2004	Single document	Unknown	Generate very short summaries (less than 75 bytes), called headlines generation. It identifies of the most relevant sentences and verb phrases were extracted using x^2 weight as threshold using genetic algorithm.	Used DUC-2003 data for implementation. Identified unigram recall but failed to identify bigrams, trigrams and four-gram results. Identified best feature combination	ROUGE-1.	Extracts

Continued on next page . . .

Table 4.1 – continued from previous page . . .

SYS, [REF], YEAR	INPUTs	DOMAIN	FEATURE	EVALUATION	METRICS	OUTPUT
Filatova, et al.,[117], 2004	Multi document	Unknown	It is a MSR-NLP Summarization system. Main goal of the system is, to explore an event-centric approach to summarization, and to explore a generation approach to summary realization. Page Rank algorithms used to identify highly weighted nodes in a document graph.	Experiment with DUC 2003, and DUC 2004 data sets. To determine quality of the generated summaries as compared with human judgments.	ROUGE	Extracts

Continued on next page . . .

Table 4.1 – continued from previous page . . .

SYS, [REF], YEAR	INPUTs	DOMAIN	FEATURE	EVALUATION	METRICS	OUTPUT
CRL/NYU ,[118], 2004	Multi document	Unknown	It is based on sentence extraction. TDT and TREC techniques are used for cluster. To estimate significance of sentences, scoring functions are used, such as position, length, tf*idf, Headline. To get redundant information, similarity between sentences is estimated. A module is used to categorize document sets into two groups corresponding to the distribution of key sentences.	To check quality question in DUC 2004.	Mean Coverage and ROUGE	Extracts

Continued on next page . . .

Table 4.1 – continued from previous page . . .

SYS, [REF], YEAR	INPUTs	DOMAIN	FEATURE	EVALUATION	METRICS	OUTPUT
CLASSY,[119], 2005	Multi document	Domain specific: news articles	It is a Query-based summarization system. ”Shallow parsing” techniques is used. To score the each sentences in a document, Hidden Markov Model is used.	To identify quality of the output summaries as compared to human made summaries. Experimented with DUC 2005 data sets.	ROUGE-1, pyramid score	Extracts
CATS,[120], 2005	Multi document	Domain specific: news articles	It is a topic-oriented Summarization system. It consists of 5 steps: Question analysis, Document analysis (TextTiling algorithm is used for thematic segmentation for each sentences), Sentence Scoring, Sentence compression, and Sentence selection.	NIST evaluates summaries in DUC 2005, to determine quality, relevance of the output summaries as compared with human made summaries.	ROUGE	Extracts

Continued on next page . . .

Table 4.1 – continued from previous page . . .

SYS, [REF], YEAR	INPUTs	DOMAIN	FEATURE	EVALUATION	METRICS	OUTPUT
ERSS,[121], 2005	Multi document	Domain specific: news articles	It is based on a single strategy, the generation and processing of conference chains using fuzzy set theory. Pipeline of processing component is run in sequence for processing documents. Important components are: POS Tagger, NE Transducer, NP/VP Chunker, Fuzzy Coreferencer.	Implemented based on the GATE framework. DUC 2004 and DUC 2005 data set is used.	ROUGE-1, ROUGE-2, ROUGE-SU4, Pyramid and Basic Element Score.	Extracts

Continued on next page . . .

Table 4.1 – continued from previous page ...

SYS, [REF], YEAR	INPUTs	DOMAIN	FEATURE	EVALUATION	METRICS	OUTPUT
MSBGA,[122], 2006	Multi document	Domain specific: news articles	Optimal summary is extracted from set of summaries formed by the conjunction of the original articles sentences. To solve NP hard optimization problem, genetic algorithm is used. To improve accuracy of term frequency, TFS method is applied.	Data set from DUC 2002, DUC 2003 and DUC 2005 is used. To check quality, performance, accuracy of summarizer.	ROUGE-1, ROUGE-W	Extracts

Continued on next page ...

Table 4.1 – continued from previous page . . .

SYS, [REF], YEAR	INPUTs	DOMAIN	FEATURE	EVALUATION	METRICS	OUTPUT
FEMsum,[123], 2007	Single and Multi document	Domain specific: news articles	Providing summary-length answers to complex questions. Based on query-focused summarization task. Uses graph to represent the relations between candidate sentences. Organized in three language independent components: Relevant Information Detector (RID), Content Extractor (CE), Summary Composer (SC).	Uses DUC-2007 data sets. Check linguistic quality against human summarizer. Two baseline s used.	Mean	Extracts

Continued on next page . . .

Table 4.1 – continued from previous page . . .

SYS, [REF], YEAR	INPUTs	DOMAIN	FEATURE	EVALUATION	METRICS	OUTPUT
QCS,[124], 2007	Single and Multi document	Unknown	It is portable, modular, and permits experimentation with different instantiations of each of the constituent text analysis components. Developed in the language C and C++ and tested under the operating systems SunOS and Linux. -developed as client server application.	DUC 2002-2004 data set are used. Experimented to measure the performance algorithm and quality of the signature terms.	ROUGE-1 and ROUGE-2.	Extracts

Continued on next page . . .

Table 4.1 – continued from previous page . . .

SYS, [REF], YEAR	INPUTs	DOMAIN	FEATURE	EVALUATION	METRICS	OUTPUT
GOFAIsum,[125], 2007	Multi document	Domain specific: news articles	Topic-answering and summarizing system developed for DUC 2007. Uses source of linguistic knowledge, FIPS. To manipulate datas, represented in a tree structures using XML and XSLT.	Evaluated by NIST. Measures responsiveness, linguistic quality of the summaries.	ROUGE-2, ROUGE-SU4.	Extracts

Continued on next page . . .

Table 4.1 – continued from previous page . . .

SYS, [REF], YEAR	INPUTs	DOMAIN	FEATURE	EVALUATION	METRICS	OUTPUT
NeTsum,[126], 2007	Multi document	Domain specific: news articles	-to extract three sentences from a single document that best match various characteristics of the three highlights. RankNet, a neural network algorithm is used to rank sentences. -to speed up the performance of RankNet is implemented in the LambdaRank	Compare against the baseline of choosing : the first three sentences as the block summary. : choosing n sentences to match highlight n.	ROUGE-1, ROUGE-2.	Extracts

Continued on next page . . .

Table 4.1 – continued from previous page . . .

SYS, [REF], YEAR	INPUTs	DOMAIN	FEATURE	EVALUATION	METRICS	OUTPUT
FastSum,[127], 2008	Multi document	Domain specific: news articles	-machine learning approach is used to rank all sentences in the topic cluster. -two sets features used: (i)Word-based : probability of words for the different container is considered. (ii) Sentence-based : length and position of the sentence in the document is considered. -regression SVM is used for learning the feature weights.	Compared the performance with DUC-2006 and DUC-2007 competitions. - Better than he PYTHY system for 2006. -evaluate the performance by applying each feature separately.	ROUGE-2.	Extracts

Continued on next page . . .

Table 4.1 – continued from previous page . . .

SYS, [REF], YEAR	INPUTs	DOMAIN	FEATURE	EVALUATION	METRICS	OUTPUT
PPRSum,[128], 2008	Multi document	Unknown	- query-based document summarization -calculate personalized view of importance of the pages. -computed global features of salience model of sentence by Nave Bayes Model. - to get salience model and relevance model of sentence in the corpus, personalized prior probability is computed. -to reduce redundancy, MMR model is used.	DUC-2007 dataset is used. To analyze the system, it compare with performance of other systems.	ROUGE-2 and ROUGE-4	Extracts

Continued on next page . . .

Table 4.1 – continued from previous page . . .

SYS, [REF], YEAR	INPUTs	DOMAIN	FEATURE	EVALUATION	METRICS	OUTPUT
Adasum,[129], 2008	Multi document	Unknown	- Adaptive model for topic-oriented text summarization system. - Summary and topic representation can be mutually boosted is assumed.	DUC-2007 data set is used. Compared with DUC 2007 top performing systems	ROUGE-2 and ROUGE-SU4	Extracts
TEXT2TABLE,[130], 2009	Multi document	Domain specific: Medical records	to identify negative event -investigate what kind of information is helpful for negative event identification. -SVM classifier is used to distinguish negative events from other events.	CRF toolkit is used for experiment. -performance in various feature combination.	Precision, Recall and F-measure	Extracts and convert into table structure

Continued on next page . . .

Table 4.1 – continued from previous page . . .

SYS, [REF], YEAR	INPUTs	DOMAIN	FEATURE	EVALUATION	METRICS	OUTPUT
OHSU,[131], 2009	Multi document	Domain Independent	Query-based summarization system -log-linear model is used to classify each word in a sentence - Sentence ranking methods are query neural ranking and query-focused ranking used.	Evaluation: DUC-2005 for training data and DUC-2006 for development data for testing different features. CSLU-OHSU1 and CSLU-CHSU2 system are used to run. -entity linking system used internal Wikipedia links. -TAC 2009 KBP query set used for evaluate performance of entity linking system.	ROUGE	Extracts

Continued on next page . . .

Table 4.1 – continued from previous page . . .

SYS, [REF], YEAR	INPUTs	DOMAIN	FEATURE	EVALUATION	METRICS	OUTPUT
Hachey,[132], 2009	Multi document	Domain specific: news article	-based on Generic Relation Extraction -model Information Extraction (IE) which includes relations -capture latent semantic similarities between connector model based on latent Dirichlet allocation -rely on dependency parsing has done.	Evaluation:DUC-2001 data sets used. Compared with the human made summaries.	ROUGE-1 and ROUGE-SU4	Extracts

Continued on next page . . .

Table 4.1 – continued from previous page . . .

SYS, [REF], YEAR	INPUTs	DOMAIN	FEATURE	EVALUATION	METRICS	OUTPUT
TIARA,[133], 2010	Multi document	Domain specific: e-mail, emergency room records.	-visual analytic system, which combines text analytics and interactive visualization to help users explore and analyze large collections of text. -used topic analysis techniques to derive topics from large documents. -Lucene is used to index each document and its associated topics. -select time-sensitive keywords for different time segments.	Evaluation:topic modeling toolkit is used to perform LDA topic analysis. Evaluate quality of time sensitive keyword selection	Precision, F-Score, completeness and distinctiveness	Extracts and visualization view

Continued on next page . . .

Table 4.1 – continued from previous page . . .

SYS, [REF], YEAR	INPUTs	DOMAIN	FEATURE	EVALUATION	METRICS	OUTPUT
Anlei et al.,[134], 2010	Multi document	Domain specific: Breaking news queries	-ranking documents by relevance which takes freshness into account. -determine query is time sensitive or not. -to rank recency sensitive queries, different categories of features are used.	Evaluation: Evaluate quality of ranking model for both online and offline experiments are done.	discounted cumulative gain (DCG) and normalized discounted cumulative gain (NDCG).	evaluation metrics
Shi et. al.,[135], 2010	Multi document	Unknown	-classifying its data facets into four categories: time facet, category facet, text content facet and associated structured facet. -navigation methods are used for manipulating, and customizing interactions of data facet. -for finding visual pattern analytical process is done.	Evaluation: To check performance of the system, two case studies are presented.	Unknown	Text Visualization

Continued on next page . . .

Table 4.1 – continued from previous page . . .

SYS, [REF], YEAR	INPUTs	DOMAIN	FEATURE	EVALUATION	METRICS	OUTPUT
MCMR,[136], 2011	Single and Multi document	Unknown	-unsupervised text summarization -optimize three properties: relevance, redundancy and length. -documents are split into sentences and select salient sentences from document (s).	DUC 2005 and DUC 2006 data sets. Measures overlap units, similarities.	ROUGE-2 and ROUGE-SU4	Extracts
Theme Crowds,[137][], 2011	Multi document	domain specific: collection of Twiiter users	-most relevant clusters of users relating to particular topic -multilevel tag cloud, convey crowd size and content compactly for a given time stamp. -Automatic Antichain Selection to specify crowd resolution.	applied to a microblogging corpus with the goal of identifying groups of users with a large geographical area studies are presented.	No	Correct Resolution and tag of summaries.

Continued on next page . . .

Table 4.1 – continued from previous page . . .

SYS, [REF], YEAR	INPUTs	DOMAIN	FEATURE	EVALUATION	METRICS	OUTPUT
SWING,[138], 2011	Multi document	Domain specific: news articles	-based on supervised machine learning approach -consists of two classes of feature: generic features and category-specific features. -calculate CSI for each sentence.	TAC-2011 data set is used. Comparison against NUS1 and NUS2.	ROUGE-2, ROUGE-SU4	Extract
Genest et al.,[139], 2011	Multi document	Domain Independent	- Concept of Information Items (INIT) is used to extract sentences. -identified all entities in the text, their properties, predicates between them, and characteristics of the predicates.	TAC 2010 data set used. Evaluate quality and overall responsiveness.	linguistic quality, pyramid score.	Abstract

Continued on next page . . .

Table 4.1 – continued from previous page . . .

SYS, [REF], YEAR	INPUTs	DOMAIN	FEATURE	EVALUATION	METRICS	OUTPUT
Tsarve et al.,[140], 2011	Multi document	Unknown	-It used both supervised and unsupervised classification techniques. -In supervised classification, multi-label is learned. -In Unsupervised classification, two clustering methods are applied such as, Flat clustering and Hierarchical clustering	DUC 2002 dataset is used. To determine quality and performance of the summary, art of summarization methods is evaluated.	ROUGE-2, ROUGE-L, Rouge-S4 and ROUGE-W.	Extract
UWN,[141], 2012	Multi document	Domain Independent	-lexical knowledge base, which describes the meanings relationships of words in over 200 languages. -automatically link terms in different languages to the meanings already defined in WordNet. -introduced MENTA and some other extensions.	Evaluate random samples of term -sense links for different languages	Precision	Extract

Continued on next page . . .

Table 4.1 – continued from previous page . . .

SYS, [REF], YEAR	INPUTs	DOMAIN	FEATURE	EVALUATION	METRICS	OUTPUT
Mirroshandel et al.,[142], 2012	Multi document	blank	-Introduced two new methods, BCDC and EMTRL for extracting temporal relations between events. -SVM is used for extracting features.	: INDRI software used for retrieve related texts. LIBSVM, for the SVM classification. EVITA, for event extraction. TDT , OTC, and TimeBank data corpus is used.	Before, After, and Overlap relations.	Extract

4.2 Experimental results on Keyword Extraction Methods

The performance of the proposed algorithm was studied on a relatively large corpus of documents. To illustrate the result, we selected the set of more than hundred articles from the set globalization articles. Each document consists of more than hundred words in average. At first the punctuations were removed from the documents. In preprocessing approach only stop word filtering is performed. To address the problem of the variable document length, adaptive window size m_i for each document was applied. Each document, K and m_i value is varied. To implement tf-idf values, idf is varied for each document. The meaningful words are extracted using two methods, one is tf-idf and the other is Helmholtz principle. Comparison of number of words extraction from above two methods is implemented. To extract meaningful words according to Helmholtz principle; expression (7) is applied from the corpus of different length documents. In Figure 4.1 when document size is increased the number of meaningful words is increasing in case of NFA. But for tf-idf, it is shown that number of meaningful words is not depend on size of documents. Number of meaningful words is more as compared with the meaningful words extracted from tf-idf in each document. We followed principle of Helmholtz Principle to calculate tf-idf. To find tf-idf, adaptive window size is applied in each document. idf and tf value is varied in each document. To separate easily the number meaningful words extracted using tf-idf with different threshold values log function is applied. $\log(tf - idf)$ value is greater than -8.5,-7.5,-6.5, compare with number of words using NFA is applied, shown in Figure 4.2. More number of meaningful words are extracted as compared with nfa. In Figure 4.3 NFA with $\log(tf - idf)$ values greater than -4.5 and -5.5 is compared. Each data is analyzed and the number of words in these documents varies dramatically. For $\log(tf-idf) > -4.5$, the number of meaningful words is greater than those in NFA up to 1500 words approximately. Beyond that the number of extracted words decreases. However, for $\log(tf-idf) > -5.5$, more number of meaningful words are extracted up to 13000 words approx then decreases.

The six most meaningful words extracted from the set of globalization articles are: economic,population, government, poor, development and political. To execute

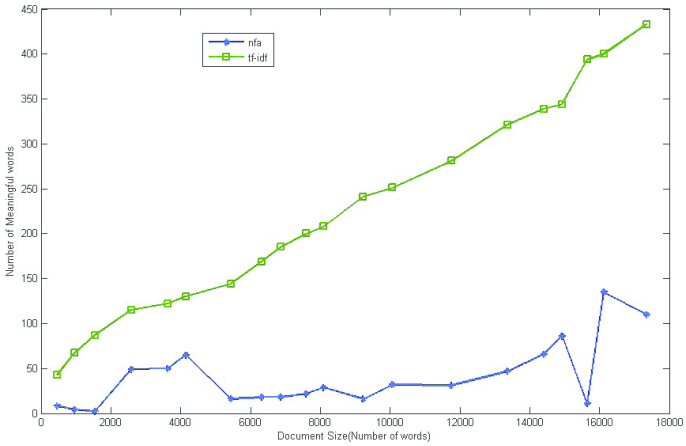


Figure 4.1:

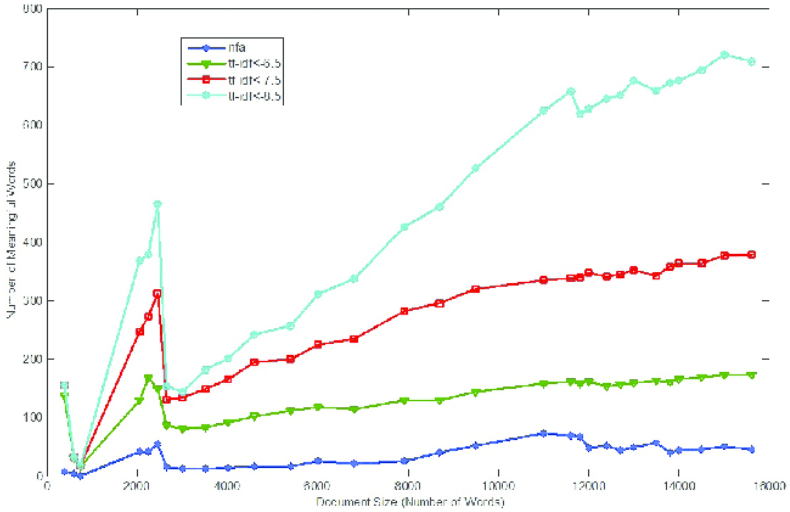


Figure 4.2:

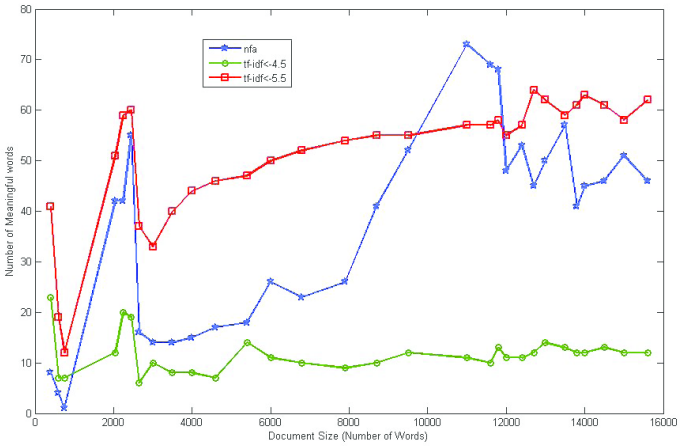


Figure 4.3:

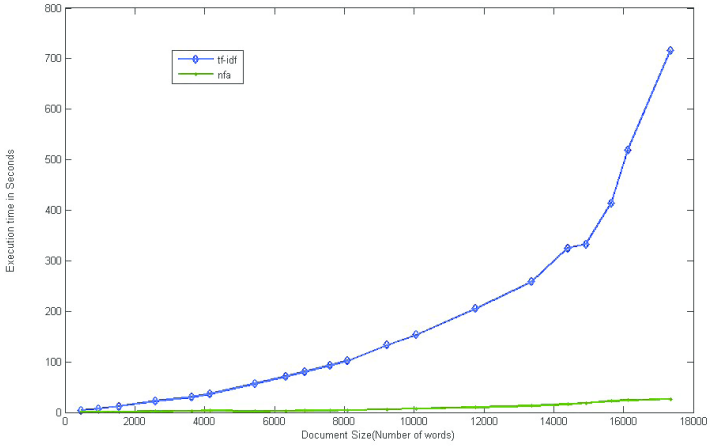


Figure 4.4:

experiment, we use python tool in high configured system. Figure 4.4 shows a comparison of run time of extracting keywords using NFA and tf-idf as executed in Figure 4. To get meaningful words using Helmholtz principle is very fast as compared to using tf-idf.

Chapter 5

Conclusion and Future work

In this thesis, we have described an overview of text summarization. We present taxonomy of text summarization based on different approaches. Categories of Evaluation methods are also explained. We have also presented a general overview of automatic text summarization systems with its main features. It is benefits for many other tasks, mainly information retrieval, information extraction or text categorization. Research on text summarization has started more than 70 years ago, still it is going on. Day by day more developed techniques are applied but still it requires improvement. In future we plan to study more systems with applied techniques which improve quality.

Keyword extraction method using Helmholtz principle was compared with the most popular Keyword extraction method i.e. tf-idf. We observe the comparison of NFA with the different level of tf-idfvalues to extract the meaningful words. Time consumed for implementing both the method to extract meaningful words was shown. When the size of documents is increased, the meaningful words are also gradually increased. Whereas for tf-idf, it is taking maximum time to implement and extracting the number of meaningful words are more as compared with NFA.

The meaningful words attained through the NFA and tf-idf method will help to create summaries of the documents. The tf-idf values can be applied in SVD to give output.We will apply evaluation measures to the output summaries from both key extraction methods and compare quality of summaries.

Bibliography

[1] J. Wang, W. Shao, and F. Zhu. Biological term boundary identification by maximum entropy model,. In *Sixth IEEE conference on Industrial Electronics and Applications*, pages 2446–2448. IEEE, June 2011.

[2] I.Mani. *Automatic Summarization*, volume 3. John Benjamins Publishing Company, 2001.

[3] E.Lloret. Text summarization : An overview. 2006.

[4] G.L.Thione, M.V.D.Berg, L.Polanyi, and C.Culy. Hybrid text summarisation:combining external relevance measures with structural analysis. In *Proceedings of the Association of Computational Linguistics, Workshop Text Summarization Branches Out*, pages 25–26, Barcelona, Spain, July 2004.

[5] M.Y.Kan. *Automatic text summarization as applied to information retrieval : Using indicative and Informative Summaries*. PhD thesis, Columbia University, 2003.

[6] L.H.Reeve, H.Han, and A.D.Brooks. The use of domain-specific concepts in biomedical text summarization. *Journal of Information Processing and Management*, 43(6):1765–1776, November 2007.

[7] F.Bex and B.Verheij. Arguments, stories and evidence: critical questions for fact-finding. In *Proceedings of Conference of the International the seventh Society for the Study of Argumentation*, pages 71–84, Sic Sat, Amsterdam, 2010.

[8] E.Sanocki L.He, A.Gupta, and J.Grudin. Automatic summarization of audio-video presentation. In *Proceeding of Seventh ACM International Conference on Multimedia*, pages 489–498, New York,USA, 1999. ACM.

[9] D.Wu and Y.Bao. A summarization of multimedia resource digital rights expression language. In *Proceedings of second International Conference on Network Security, Wireless Communications and Trusted Computing*, pages 374–377. IEEE, 2010.

[10] E.B.Euripides, E.G.M.Petrakis, and E.Milios. Automatic website summarization by image content: A case study with logo and trademark images. *IEEE Transaction on Knowledge and Data Engineering*, 20(9):1195–1204, September 2008.

[11] G.Manson and S.A.Berrani. Automatic tv broadcast structuring. *International Journal of Digital Multimedia Broadcasting*, pages 153160–153176, January 2010.

[12] **T.S.Guzella and W.M.Caminhas. A review of machine learning approaches to spam filtering,. *Journal of Expert Systems with Applications,Elsevier*, 36(7):10206–10222, September 2009.**

[13] **J.C.Gomez, E.Boiy, and M.F.Moens. Highly discriminative statistical features for e-mail classification. *Journal of Knowledge and Information Systems,Springer*, 31(1):23–53, April 2012.**

[14] **X.L.Wang, H.Zhao, and B.L.Lu. Automated quality assessment of web pages form textual content. In *Proceedings of the International Conference on Machine Learning and Cybernetics*, pages 2000–2006. IEEE, July 2012.**

[15] **W.Hersh. *Informative Retrieval: a Health and Biomedical Perspective*. Springer, New York, 2009.**

[16] **S.Teufel and M.Moens. Summarizing scientific articles: Experiments with relevance and rhetorical status. *Journal of Computational Linguistics*, 28(4):409–445, December 2002.**

[17] **I.T.Medeni, S.Peker, and M.E.Uyar. A knowledge visulization model for evaluating internet news agencies on conflicting news. In *Proceedings of the thirty fourth International Convention, MIPRO*, pages 850–853. IEEE, 2011.**

[18] **H.P.Luhn. The automatic creation of literature abstract. *IBM Journal of Research and Development*, 2(2):159–165, April 1958.**

[19] **A.Ramanathan and D.D.Rao. A lightweight stemmer for hindi. In *In Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics, on Computatinal Linguistics for South Asian Languages Workshop*, April 2003.**

[20] **U.MIshra and C.Prakash. Maulik: An effective stemmer for hindi language. *International Journal on Computer Science and Engineering*, 4(5):711–717, May 2012.**

[21] **H.P.Edmundson. New methods in automatic extracting. *Journal of Association of Computational Linguistics*, 16(2):264–285, April 1969.**

[22] **M.Chandra, V.Gupta, and S.Paul. A statistical approach for automatic text summarization by extraction. In *Proceeding of the International Conference on Communication Systems and Network Technologies*, pages 268–271. IEEE, 2011.**

[23] **M.R.Murthy, J.V.R.Reddy, P.P.Reddy, and S.C.Satapathy. Statistical approach based keyword extraction aid dimensionality reduction. In *Proceedings of the International Conference on Information Systems Design and Intelligent Applications*. Springer, 2011.**

[24] **J.Morris and G.Hirst. Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Journal of Computational Linguistics, ACM*, 17(1):21–48, March 1999.**

[25] **T.Pedersen, S.Patwardhan, and J.Michelizzi. Wordnet::similarity : Measuring the relatedness of concepts. In *Proceedings in Human Language Technology Conference - North American chapter of the Association for Computational Linguistics Annual Meeting- Demonstrations*, pages 38–41. ACM, 2004.**

[26] **H.G.Silber and K.F.McCoy. Efficient text summarization using lexical chains. In *Proceedings of Fift International Conference on Intelligent User Interfaces*, pages 252–255, New York, USA, 2000. ACM.**

[27] **G.Ercan and I.Cicekli. Using lexical chains for keyword extraction. *International Journal of Information Processing and Management,ACM*, 43(6):1705–1714, November 2007.**

[28] **W.C.Mann and S.A.Thompson. Rhetorical structure theory: Towards a function theory of text organization. *Text - Interdisciplinary Journal for the Study of Discourse*, 8(3):243–281, November 1988.**

[29] **V.R.Uzeda, T.Pard, and M.Nunes. Evaluation of automatic summarization methods based on rhetorical structure theory. In *Eighth International Conference on Intelligent Systems Design and Applications*, pages 389–394. IEEE, November 2008.**

[30] **L.Chengcheng. Automatic text summarization based on rhetorical structure theory. In *International Conference on Computer Application and System Modeling*, pages 595–598, Piscataway, NJ, USA, October 2010. IEEE.**

[31] **M.Litvak and M.Last. Graph based keyword extraction for single document summarization. In *Proceedings of the Workshop on Multi-source Multilingual, Information Extraction and Summarization*, pages 17–24. ACM, 2008.**

[32] **S.Brin and L.Page. The anatomy of a large-scale hypertextual web search engine. In *Proceedings of the Seventh International World Wide Web Conference, Computer Networks and ISDN Systems*, pages 107–117. Elsevier, April 1988.**

[33] **R.Mihalcea. Graph-based ranking algorithms for sentence extraction, applied to text summarization. In *Proceedings of the Association for Computational Lingusitics on Interactive poster and demonstration sessions*. ACM, 2004.**

[34] **D.Horowitz and S.D.Kamvar. Anatomy of a large-scale social search engine. In *Nineteenth International Conference on World Wide Web*, pages 431–440, New York, USA, 2010. ACM.**

[35] **M. Efron. Information search and retrieval in microblogs. *Journal of the American Society for Infromation Science and Technology*, 62(6):996–1008, June 2011.**

[36] **D.D.Lewis. Naive (bayes) at forty: The independence assumption in information retrieval. In *Proceedings of tenth Europian Conference on Machine Learning*, pages 4–15, London, UK, 1998. Springer- Verlag.**

[37] **J.D.M.Rennie, L.Shih, J.Teevan, and D.R.Karger. Tackling the poor assumptions of nave bayes classifiers. In *Proceedings of International Conference on Machine Learning*, pages 616–623, 2003.**

[38] **T.Mouratis and S.Kotsiantis. Increasing the accuracy of discriminative of multinomial bayesian classifier in text classification. In *Proceedings of fourth International Conference on Computer Sciences and Convergence Information Technology*, pages 1246–1251. IEEE, November 2009.**

[39] **C.Y.Lin. Training a selection function for extraction. In *Proceedings of the eighth International Conference on Information and Knowledge Management*, pages 55–62, NY, USA, 1999. ACM.**

[40] **K.Knight and D.Marcu. Statistics-based summarization- step one: Sentence compression. In *Proceeding of the Seventeenth National Conference of the American Association for Artificial Intelligence*, pages 703–710, 2000.**

[41] **L.Rabiner and B.Juang. An introduction to hidden markov model. *Acoustics Speech and Signal Processing Magazine,*, 3(1):4–16, April 2003.**

[42] **R.Nag, K.H.Wong, and F.Fallside. Script recognition using hidden markov models. In *Proceedings of International Conference on Acoustics Speech and Signal Processing*, pages 2071–2074. IEEE, April 1986.**

[43] **C.Zhou and S.Li. Research of information extraction algorithm based on hidden markov model. In *Proceedings of second International Conference on Information Science and Engineering*, pages 1–4. IEEE, December 2010.**

[44] **M.Osborne. Using maximum entropy for sentence extraction. In *Proceedings of the AssociaCL-02 Workshop on Automatic Summarization*, pages 1–8, Morristown, NJ, USA, 2002. Association for Computational Linguistics.**

[45] **B.Pang, L.Lee, and S.Vaithyanathan. Thumbs up? : Sentiment classification using machine learning technique. In *Proceedings of the Association for Computational Linguistics conference on Empirical methods in Natural Language Processing*, pages 79–86. ACM, 2002.**

[46] **H.L.Chieu and H.T.Ng. A maximum entropy approach information extraction from semi-structured and free text. In *Proceedings of the Eighteenth National Conference on Artificial Intelligence, American Association for Artificial Intelligence*, pages 786–791, Menlo Park, Ca, USA, 2002.**

[47] **R.Malouf. A comparison of algorithms for maximum entropy parameter estimation. In *Proceedings of sixth conference on Natural Language Learning*, pages 1–7, Stoudsburg, PA, USA, 2002. ACM.**

[48] **P.Yu, J.Xu, G.L.Zhang, Y.C.Chang, and F.Seide. A hidden−state maximum entropy model forward confidence estimation. In *International Conference on Acoustic, Speech and Signal Processing*, pages 785–788. IEEE, 2007.**

[49] **M.E.Ruiz and P.Srinivasan. Automatic text categorization using neural networks. In *Proceedings of the Eighth American Society for Information Science/ Classification Research, American Society for Information Science*, pages 59–72, Washington, 1997.**

[50] **T.Jo. Ntc (neural network categorizer) neural network for text categorization. *International Journal of Information Studies*, 2(2), April 2010.**

[51] **P.M.Ciarelli, E.Oliveira, C.Badue, and A.F.De-Souza. Multi-label text categorization using a probabilistic neural network. *International Journal of Computer Information Systems and Industrial Management Applications*, 1:133–144, 2009.**

[52] **I.Guyon, J.Weston, S.Barnhill, and V.Vapnik. Gene selection for cancer classification using support vector machines. mach. learn. *Machine Learning*, 46(1–3):389–422, March 2002.**

[53] **T.Hirao, Y.Sasaki, H.Isozaki, and E.Maeda. Ntt's text summarization system for duc-2002. In *Proceedings of the Document Uderstanding Conference*, pages 104–107, 2002.**

[54] **L.N.Minh, A.Shimazu, H.P.Xuan, B.H.Tu, and S.Horiguchi. Sentence extraction with support vector machine ensemble. In *Proceedings of the First World Congress of the International Federation for Systems Research*, pages 14–17, Kobe, Japan, November 2005.**

[55] **T.K.Landauer, P.W.Foltz, and D.Laham. Introduction to latent semantic analysis. *Journal of Discourse Processes,*, 25(2–3):259–284, 1998.**

[56] **B.Rosario. Latent semantic indexing: An overview. Technical report, Technical Report of INFOSYS 240, University of California, 2000.**

[57] **T.Hofmann. Unsupervised learning by probabilistic latent semantic analysis. *Journal of Machine Learning*, 42(1–2):177–196, January–February 2001.**

[58] **M.Simina and C.Barbu. Meta latent semantic analysis. In *IEEE International conference on Systems, Man and Cybernetics*, pages 3720–3724. IEEE, October 2004.**

[59] **J.T.Chien. Adaptive bayesian latent semantic analysis. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(1):198–207, January 2008.**

[60] **D.Wang, T.Li, S.Zhu, and C.Ding. Multi-document summarization via sentence-level semantic analysis and symmetric matrix factorization. In *Proceedings of the Thirty First Annual International ACM Special Interest Group on Information Retrieval Conference on Research and Development in Information Retrieval*, pages 307–314, New York, USA, 2008. ACM.**

[61] **J.H.Lee, S.Park, C.M.Ahn, and D.Kim. Automatic generic document summarization based on non-negative matrix factorization. *Journal on Information Processing and Management*, 45(1):20–34, January 2009.**

[62] **T.G.Kolda and D.P.O'Leary. Algorithm 805: Computation and uses of the semidiscrete matrix decomposition. *Transactions on Mathematical Software*, 26(3):415–435, September 2000.**

[63] **V.Snasel, P.Moravec, and J.Pokorny. Using semi-discrete decomposition for topic identification. In *Proceedings of the Eighth International Conference on Intelligent Systems Design and Applications*, pages 415–420, Washington, DC, USA, 2008. IEEE.**

[64] **D.R.Radev, E.Hovy, and K.McKeown. Introduction to the special issue on summarization. *Journal of Computational Linguistics*, 28(4):399–408, December 2002.**

[65] **J.S.Kallimani, K.G.Srinivasa, and B.E.Reddy. Information retrieval by text summarization for an indian regional language. In *Proceedings of IEEE Natural Language Processing and Knowledge Engineering*, pages 1–4, Beijing, China, August 2010.**

[66] **M.J.Witbrock and V.O.Mittal. Ultra-summarization: A statistical approach to generating highly condensed non-extractive summaries. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 315–316, New York, NY, USA, 1999. ACM.**

[67] **I.Mani, D.House, G.Klein, L.Hirschman, T.Firmin, and B.Sundhein. The tipster summac text summarization evaluation. In *Proceedings of the Ninth Conference on European chapter of the Association for Computational Linguistics*, pages 77–85. ACM, 1999.**

[68] **C.Y.Lin and E.Hovy. Manual and automatic evaluation of summaries. In *Proceedings of the Association Computational Linguistics-02 Workshop on Automatic Summarization*, pages 45–51. ACM, 2002.**

[69] **K.Papineni, S.Roukos, T.Ward, and W.J.Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the fortieth Annual Meeting on Association for Computational Lingusitics*, pages 311–318. ACM, July 2002.**

[70] **M.Hassel. *Evaluation of Automatic Text Summarization*. PhD thesis, School of Computer Science and Communication,Royal Institute of Technology, Stockholm, Sweden, 2004.**

[71] **J.Steinberger and K.Jezek. Evaluation measures for text summarization. *Computing and Informatics*, 28(2):251–275, 2009.**

[72] **C.Buckley and E.M.Voorhees. Evaluating evaluation measure stability. In *Proceedings of the Twenty Third Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 33–40, NY, USA, 2000. ACM.**

[73] **D.R.Radev and D.Tam. Single-document and multi-document summary evaluation via relative utility. In *Proceedings of the ACM Conference on Information and Knowledge Management*, New Orleans, LA, 2003. ACM.**

[74] **H.Saggion, D.Radev, S.Teufel, W.Lam, and S.M.Strassel. Developing infrastructure for the evaluation of single and multidocument summarization systems in a cross-lingual environment. In *In Proceedings of Third International Conference on Language Resources and Evaluation*, pages 747–754, Las Palmas, Gran Canaria , Spain, 2002.**

[75] **C.Y.Lin. Rouge: A package for automatic evaluation of summaries. In *Proceedings of the Association for Computational Linguistics Workshop on Text Summarization Branches Out*, pages 74–81, 2004.**

[76] **S.Maskey and A.Rosenberg. Power mean pyramid scores for summarization evaluation. In *Thirteenth Annual Conference of the International Speech Communication Association*, Portland, Oregon, USA, September 2002.**

[77] **J.Kaur and V.Gupta. Effective approaches for extraction of keywords. *International Journal of Computer Science*, 7(6), November 2010.**

[78] **E.Frank, G.W. Paynter, I.H.Witten, C.Gutwin, and C.G.Nevill. Domain specific key phrase extarction. In *Proceeding of Sixteenth International Conference on Artificial Intelligence*, pages 668–673, San Francisco, CA,USA, 1999.**

[79] **I.H.Witten, G.W. Paynter, E.Frank, C.Gutwin, and N. Manning C.G. Kea: Practical automatic keyphrase extraction. In *Proceeding of fourth ACM conference on Digital libraries*, pages 254–=255, New York, USA, 1999. ACM.**

[80] **P.D.Turney. Learning algorithm for keyphrase extraction. *Journal of Information Retrieval*, 2(4):303–336, May 1999.**

[81] **Y.H. Kerner, Z. Gross, and A. Masa. Automatic extraction and learning of keyphrases from scientific articles. In *Proceedings of sixth International Conference on Computational Linguistics and Intelligent Text Processing*, pages 657–669, Mexico City , Mexico, February 2005. Springer.**

[82] **F. Liu, D. Pennell, F. Liu, and Y. Liu. Unsupervised approaches for automatic keyword extraction using meeting transcripts. In *Proceedings of Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 620–628, Stroudsburg, PA, USA,, June 2009.**

[83] **K. Barker and N. Cornacchia. Using noun phrase heads to extract document keyphrases. In *Proceeding of thirteenth Biennial Conference of the Canadian Society on Computational Studies of Intelligence*, pages 40–52, Berlin, German, May 2000. Springer.**

[84] **E. Gaussier B. Daille and J.M.Lange. Towards automatic extraction of monolingual; and bilingual terminology. In *Proceeding of the fifteenth conference on Computational linguistics*, pages 515–521, Stroudsburg, PA, USA, 1994. Springer.**

[85] **J.E.Burnett, D.Cooper, M.F.Lynch, P.Willett, and M.Wycherley. Document retrieval experiments using indexing vocabularies of varying size. 1.variety generation symbols assigned to the fronts of index terms. *Journal of Documentation*, 35:197–206, 1979.**

[86] **J.D.Cohen. Highlights: Language- and domain-independent automatic indexing terms for abstracting. *Journal of the American Society for Information Science*, 46(3):162–174, April 1995.**

[87] **Y.Matsuo and M.Ishizuka. Keyword extraction from a single document using word co-occurrence statistical information. *International Journal on Artificial Intelligence Tools*, 13(1):157–169, 2004.**

[88] **H.G.Silber and K.F.McCoy. Efficient text summarization using lexical chains. In *Proceedings of Fifth International Conference on Intelligent User Interfaces,*, pages 252–255, New York, USA, January 2000. ACM.**

[89] **A.Hulth. Improved automatic keyword extraction given more linguistic knowledge. In *Proceeding of Conference on Empirical Methods in natural language processing,*, pages 216–223, Stroudsburg, PA, USA, 2003.**

[90] **R.Brazilay and M.Elhadad. Using lexical chains for text summarization. In *Proceeding of the ACL'97/EACL workshop on Intelligent Scalable Text Summarization*, pages 10–17, Madrid, Spain, July 2007.**

[91] **R.Angheluta, R.Busser, and M.Moens. The use of topic segmentation for automatic summarization. In *Proceedings of the workshop on automatic summarization*, pages 66–70, Philadelphia,PA, USA, 2002.**

[92] **Y.Uzun. Keyword extraction using nave bayes. 2005.**

[93] **K.Zhang, H.Xu, J.Tang, and J.Li. Keyword extraction using support vector machine. In *Proceedings of seventh International Conference on Web-Age Information Management*, pages 85–96, Springer-Verlag, Berlin, Germany, June 2006.**

[94] **F.Liu, F.liu, and Y.Liu. Automatic keyword extraction for the meeting corpus using supervised approach and bigram expansion. In *IEEE Workshop on Spoken Language Technology*, pages 181–184, NJ, USA, December 2008. IEEE.**

[95] **J.B.K. Humphreys. Phraserate: An html keyphrase extractor. Technical report, University of California, Riverside, California,, 2002.**

[96] **J.J.Pollock and A.Zamora. Automatic abstracting research at chemical abstracts service. *Journal of Chemical Information and Computer Sciences*, 15(4):226–232, 1975.**

[97] **R.Brandow, K.Mitze, and L.F.Rau. Automatic condensation of electronic publications by sentence selection. *Journal of Information Processing and Management*, 31(5):675–685, September 1995.**

[98] **C.Aone, M.E.Okurowski, and J.Gorlinsky. Trainable, scalable summarization using robust nlp and machine learning. In *Seventeeth International Conference on Computational Linguistics*, pages 62–66, 1998.**

[99] **D.R.Radev and K.R.McKeown. Generating natural language summaries from multiple on-line sources. *Journal of Computer Linguistics : Special issue on natural language generation*, 24(3):470–500, September 1998.**

[100] **E.Hovy and C.Y.Lin. Automated text summarization and the summarist system. In *Proceedings of a workshop on held at Baltimore*, pages 13–15, Baltimore, Maryland, October 1998.**

[101] **D.Marcu. Discourse trees are good indicators of importance in text. In *Advances in Automatic Text Summarization*, pages 123–136. The MIT Press, 1999.**

[102] **R.Brazilay, K.R.McKeown, and M.Elhadad. Information fusion in the context of multi-document summarization. In *Proceedings of the thiry seventh Annual Meeting of the Association Computational Linguistics on Computational Linguistics*, pages 550–557. ACM, 1999.**

[103] **H.H.Chen and C.J.Lin. A multilingual news summarizer. In *Proceedings of the Eighteenth conference on Computational linguistics*, pages 159–165. ACL, 2000.**

[104] D.R.Radev. Experiments in single and multidocument summarization using mead. In *Proceedings of First Document Understanding Conference*, New Orleans, LA, 2001.

[105] D.R.Radev, S.Blair-Goldensohn, Z.Zhang, and R.S.Raghavan. Newsessence: A system for domain-independent, real-time news clustering and multi-document summarization. In *Proceedings of the First International Conference on Human Language Technology Research*, pages 1–4. ACM, 2001.

[106] D.R.Radev, S.Blair-goldensohn, and Z.Zhang. Webinessence: a personalized web-based multi-document summarization and recommendation system. In *North American Chapter of the Association for Computational Linguistics: Human Language Technologies Workshop on Automatic Summarization*, pages 79–88, 2001.

[107] C.Y.Lin and E.Hovy. Automated multi-document summarization in neats. In *Proceedings of the second international conference on Human Language Technology Research*, pages 59–62, San Francisco, CA, USA, 2002. Morgan Kaufmann Publishers Inc.

[108] K.R.McKeown, R.Barzilay, D.Evans, and V.Hatzivassiloglou. Tracking and summarizing news on a daily basis with columbia's newsblaster. In *Proceedings of the Second International Conference on Human Language Technology Research*, pages 280–285, 2002.

[109] H.Daume III, A.Echihabi, D.Marcu, D.S.Munteanu, and R.Soricut. Gleans: A generator of logical extracts and abstracts for nice summaries. In *Proceedings of the Workshop on Automatic Summarization*, pages 9–14, Philadelphia, PA, 2002.

[110] S.H.Finely and S.M.Harabagiu. Generating single and multi-document summaries with gistexter. In *Proceedings of the Workshop on Automatic Summarization*, pages 30–38, Gaithersburg, MD, July 2002. NIST.

[111] H.Saggion and G.Lapalme. Generating indicative-informative summaries with sumum. *Journal of Computational Linguistics*, 28(4):497–526, December 2002.

[112] H.Saggion, K. Bontcheva, and H. Cunningham. Robust generic and query-based summarization. In *Proceedings of the tenth Conference on European chapter of the Association for Computational Linguistics*, pages 235–238, 2003.

[113] M.Brunn, Y.Chali, and B.Dufour. The university of lethbridge text summarizer at duc 2003. In *Proceedings of the text summarization workshop and 2003 document understanding conference*, pages 148–152, 2002.

[114] T.Copeck and S.Szpakowicz. Picking phrases, picking sentences. In *Proceedings of the Workshop on Automatic Summarization, Human Language Technology/ North America Chapter of the Association for Computational Linguistics*, 2003.

[115] G.Erkan and D.R.Radev. The university of michigan at duc 2004. In *Proceedings of the Document Understanding Conferences*, pages 120–127, 2004.

[116] **E.Alfonseca, A.Moreno-Sandoval, and J.M.Guirao. Description of the uam system for generation very short summaries at duc 200. In *Proceedings of the Human Language Technology/ North America Chapter of the Association for Computational Linguistics workshop on Automatic Summarization/Document Understanding Conference*. ACL, 2004.**

[117] **E. Filatova. Event-based extractive summarization. In *Proceedings of Association Computational Linguistics Workshop on Summarization*, pages 104–111, 2004.**

[118] **C.Nobata and S.Sekine. Crl/nyu summarization system at duc-2004. In *Proceedings of the Document Inderstanding Conference (DUC)*, 2004.**

[119] **J.M.Conoroy and J.D.Schlesinger. Classy query-based multi-document summarization,. In *Proceedings of the Document Understanding Workshop at the Human Language Technology Conference/Conference on Empirical Methods in Natural Language Processing*, Boston, 2005.**

[120] **A.Farzindar, F.Rozon, and G.Lapalme. Cats a topic-oriented multi-document summarization system at duc 2005. In *Proceedings of Document Understanding Workshop,*, 2005.**

[121] **R.Witte, R.Kreste, and S.Bergler. Erss 2005: Coreference-based summarization reloaded. In *Proceedings of Document Understanding Conference Workshop at the Human Language Technology Conference/Conference on Empirical Methods in Natural Language Processing*, pages 9–10, Vancouver, B.C., Canada, October 2005.**

[122] **Y.X.He, D.X.Liu, D.H.Ji, H.Yang, and C.Teng. Msbga: A multi-document summarization system based on genetic algorithm. In *Proceedings of the Fifth International Conference on Machine Learning and Cybernetics*, pages 2659–2664, August 2006.**

[123] **M.Fuentes, H.Rodr'quez, and D.Ferres. Femsum at duc 2007. In *Proceedings of the Document Understanding Conference at Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, Rochester, NY, 2007.**

[124] **D.M.Dunlavy, D.P.O'Leary, J.M.Conroy, and J.D.Schlesinger. Qcs: A system for querying, clustering and summarizing documents. *Journal of Information Processing and Management*, 43(6):1588–1605, 2007.**

[125] **F.Gotti, G.Lapalme, M.Quebec, L.Nerima, and E.Wehrli. Gofaisum: a symbolic summarizer for duc. In *Proceedings of Document Understanding Conference*, 2007.**

[126] **K.M.Svore, L.Vanderwende, and C.J.C.Burges. Enhancing single-document summarization by combining ranknet and third-party sources. In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Procssing and Computational Natural Language Learning*, pages 448–457, 2007.**

[127] **F.Schilder and R.Kondadadi. Fastsum: Fast and accurate query-based multi-document summarization. In *Proceedings of forty sixth Annual Meeting of the Association for Computational Linguistics on Human Language Technologies*, pages 205–208. ACM, 2008.**

[128] Y.Liu, X.Wang, J.Zhang, and H.Xu. Personalized pagerank based multi-document summarization. In *Proceedings of the IEEE International Workshop on Semantic Computing and Systems*, pages 169–173, Washington, DC, USA, 2008. IEEE.

[129] J.Zhang and X.Cheng. Adasum: An adaptive model for summarization. In *Proceeding of Conference on Information and Knowledge Management,*, pages 26–30, California, USA, October 2008. ACM.

[130] E.Aramaki, Y.Miura, M.Tonoike, T.Ohkuma, H.Mashuichi, and K.Ohe. Text2table: medical text summarization system based on named entity recognition and modality identification. In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing*, pages 185–192. ACL, 2009.

[131] S.Fisher, A.Dunlop, B.Roark, Y.Chen, and J.Burmeister. Ohsu summarization and entity linking systems. In *Proceedings of the Second Text Analysis Conference*, Gaithersburg, November 2009. NIST.

[132] B.Hachey. Multi-document summarisation using generic relation extraction. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 420–429, Stoudsburg, PA, USA, 2009. ACL.

[133] F.Wei, S.Liu, Y.Song, S.Pan, M.X.Zhou, W.Qian, L.Shi, L.Tan, and Q.Zhang. Tiara: a visual exploratory text analytic system. In *Proceedings of the Sixteenth Association for Computing Machinery's Special Interest Group on Knowledge Discovery and Data Mining, International Conference on Knowledge discovery and data mining*, pages 153–162, New York, NY, USA, 2010. ACM.

[134] A.Dong, Y.Chang, Z.Zheng, G.Mishne, J.Bai, R.Zhang, K.Buchner, C.Liao, and F.Diaz. Towards recency ranking in web search. In *Proceedings of the Third ACM international conference on Web search and data mining*, pages 11–20, 2010.

[135] L.Shi, F.Wei, S.Liu, L.Tan, X.Lian, and M.X.Zhou. Understanding text corpora with multiple facets. In *IEEE Symposium on Visual Analytics Science and Technology*, pages 99–106, 2010.

[136] R.M.Alguliev, R.M.Aliguliyev, M.S.Hajirahimova, and C.A.Mehdiyev. Mcmr: Maximum coverage and minimum redundant text summarization model. *Journal on Expert Systems with Applications*, 38(12):14514–14522, 2011.

[137] D.Archambault, D.Greene, P.Cunningham, and N.J.Hurley. Themecrowds: multiresolution summaries of twitter usage. In *Proceedings of the 3rd international workshop on Search and mining user-generated contents*, pages 77–84, New York, NY, USA, 2011. ACM.

[138] J.P.Ng, P.Bysani, Z.Lin, M.Y.Kan, and C.L.Tan. Exploiting category-specific information for multi-document summarization. In *Proceeding of the Text Analysis Conference*, pages 2093–2108, Gaithersburg, Maryland, USA, November 2011.

[139] P.E.Genest and G.Lapalme. Framework for abstractive summarization using text-to-text generation. In *Proceedings of the Workshop on Monolingual Text-To-Text Generation. Association for Computational Linguistics*, 2011.

[140] D.Tsarev, M.Petrovskiy, and I.Mashechkin. Using nmf-based text summarization to improve supervised and unsupervised classification. In *Eleventh International Conference on Hybrid Intelligent Systems*, pages 185–189. IEEE, December 2011.

[141] G.de Melo and G.Weikum. Uwn: A large multilingual lexical knowledge base. In *Proceedings of the Association for Computational Linguistics 2012 System Demonstrations*, pages 151–156. ACL, 2012.

[142] S.A.Mirroshandel and G.Gassem-Sani. Towards unsupervised learning of temporal relations between events. *Journal of Artificial Intelligence Research*, 45(1):125–163, September 2012.