Improving Text Summarization using Fuzzy Logic & Latent Semantic Analysis

Mr.S.A.Babar
Computer Science & Engineering.
Rajarambapu Institute of Technology,
Sakharale, India
samrat.babar@ritindia.edu

Prof.S.A.Thorat
Computer Science & Engineering.
Rajarambapu Institute of Technology,
Sakharale, India
sandip.thorat@ritindia.edu

Abstract—In this new generation, where the tremendous information is available on the internet, it is difficult to extract the information quickly and most efficiently. There are so many text materials available on the internet, in order to extract the most relevant information from it, we need a good mechanism .This problem is solved by the Automatic Text Summarization mechanism. "Text Summarization" is a process of creating a shorter version of original text that contains the important information. Text summarization can be broadly classified into two types: Extraction and Abstraction. This paper focuses on the Fuzzy logic Extraction approach for text summarization and the semantic approach of text summarization using Latent Semantic Analysis.

Keywords-Text summarization; Fuzzy logic; fuzzy rule;LSA.

I. INTRODUCTION

Before going to the Text summarization, first we, have to know that what a summary is. A summary is a text that is produced from one or more texts, that conveys important information in the original text, and it is of a shorter form [1]. The goal of automatic text summarization is presenting the source text into a shorter version with semantics [2]. The most important advantage of using a summary is, it reduces the reading time. Text Summarization methods can be classified into extractive and abstractive summarization [1]. An extractive summarization method consists of selecting important sentences, paragraphs etc. from the original document and concatenating them into shorter form. An Abstractive summarization is an understanding of the main concepts in a document and then expresses those concepts in clear natural language.

There are two different groups of text summarization [2]: Indicative and Informative. Inductive summarization only represents the main idea of the text to the user. The typical length of this type of summarization is 5 to 10 percent of the main text. On the other hand, the informative summarization systems gives concise information of the main text .The length of informative summary is 20 to 30 percent of the main text.

The automatic summarization means an automatically summarized output is given when an input is applied. Remember that input is well structured document. For this there are initially preprocesses such as Sentence Segmentation, Tokenization, Removing stop words and Word Stemming. Sentence Segmentation is separating document into sentences. Tokenization means separating sentences into words. Removing stop words means removing frequently occurring words such as a, an, the etc. And word stemming means removing suffixes and prefixes. After preprocessing each sentence is represented by attribute of vector of features. For each sentence there are 8 features and each feature has a value between 0 and 1. The 8 features are: Title features, Sentence length, Term weight, Sentence position, Sentence to sentence similarity, Proper noun, thematic word and Numerical data. Our approach is as follows: After extraction of 8 features the result is passed to fuzzifier then to inference engine and finally to defuzzifier. Rules for Inference engine is supplied from Fuzzy rule base. After this each sentence will have score and the sentence is sorted in the decreasing order of the score. Then 20% of these finally sorted sentences will be the summary of the given document.

II. RELATED WORKS

The first Automatic text summarization was created by Luhn in 1958[1] based on term frequency. Automatic text summarization system in 1969, which, in addition to the standard keyword method (i.e., frequency depending weights), also used the following three methods for determining the sentence weights: a) Cue Method b) Title Method c) Location Method. The Trainable Document Summarizer in 1995 performs sentence extracting task, based on a number of weighting heuristics. Following features were used and evaluated [2]:

1. Sentence Length Cut-O Feature: sentences containing less than a pre-specified number of words are not included in the abstract
2. Fixed-Phrase Feature: sentences containing certain cue words and phrases are included
3. Paragraph Feature: this is basically equivalent to Location Method feature
4. Thematic Word Feature: the most frequent words are defined as thematic words. Sentence scores are functions of the thematic words' frequencies
5. Uppercase Word Feature: upper-case words (with certain obvious exceptions) are treated as thematic words.

In 1990s the machine learning techniques in Natural Language Processing used statistical techniques to produce document summaries. They have used a combination of appropriate features and learning algorithms. Other approaches have used hidden Markov models and log-linear models to improve extractive summarization. Now a day's neural networks are used to generate summary for single documents using extraction. Ladda Suanmali [4] in his work has used sentence weight, a numerical measure assigned to each sentence and then selecting sentences in descending order of their sentence weight for the summary. Recently, neural networks are used to generate summary for single documents using extraction [6].

A lot of work has been done in single document and multi document summarization using statistical methods. A lot of researchers are trying to apply this technology to a variety of new and challenging areas, including multilingual summarization and multimedia news broadcast.

III. MOTIVATION FOR TEXT SUMMARIZATION

Text Summarization is an active field of research in both the IR and NLP communities.

- People keep up with the world affairs by listening to news bites.
- People even go to movies largely on the basis of reviews they've seen.
- People base investment decisions on stock market updates.
- With summaries, People can make effective decisions in less time.
- The motivation here is to build such tool which is computationally efficient and creates summaries automatically.

IV. APPROACHES TO SUMMARIZATION

Text summarization approach [5] consists of following stages:

A. Preprocessing
B. Feature Extraction
C. Fuzzy logic scoring
D. Sentence selection and Assembly

A. Preprocessing

There are four steps in preprocessing:

1. Segmentation: It is a process of dividing a given document into sentences.
2. Removal of Stop words: Stop words are frequently occurring words such as 'a' an', the' that provides less meaning and contains noise. The Stop words are predefined and stored in an array.
3. Tokenization:
4. Word Stemming: converts every word into its root form by removing its prefix and suffix so that it can be used for comparison with other words.

B. Feature Extraction

The text document is represented by set, D= {S1, S2, - - - , Sk} where, Si signifies a sentence contained in the document D. The document is subjected to feature extraction. The important word and sentence features to be used are decided .This work uses features such as Title word, Sentence length, Sentence position, numerical data, Term weight, sentence similarity, existence of Thematic words and proper Nouns .

1. Title word: A high score is given to the sentence if it contains words occurring in the title as the main content of the document is expressed via the title words. This feature is computed as follows:
   F1=Nt/Ntotal
2. Sentence Length: Eliminate the sentences which are too short such as datelines or author names. For every sentence the normalized length of sentence is calculated as:

$$F2 = \frac{\text{Number of words in the sentence}}{\text{Number of words in the longest sentence}}$$

3. Sentence Position: The sentences occurring first in the paragraph have highest score. Suppose a paragraph has n sentences then the score of every sentence for this feature is calculated as follows:

$$F3(S_1) = n/n; \quad F3(S_2)=4/5; \quad F3(S_3)=3/5; \quad F3(S_4)=2/5; \quad \text{and so on.}$$

4. Numerical data: The sentences having numerical data can reflect important statistics of the document and may be selected for summary. Its score is calculated as:

$$F4(Si) = \frac{\text{Number of numerical data in the sentence } Si}{\text{Sentence Length}}$$

5. Thematic words: These are domain specific words with maximum possible relativity. The score for this feature is calculated as the ratio of the number of thematic words that occurs in a sentence over the maximum number of thematic words in a sentence.

$$F5(Si) = \frac{\text{Number of Thematic data in the sentence } Si}{\text{Max no of thenatic words}}$$

6. Sentence to Sentence Similarity: For each sentence S, the similarity between S and every other sentence is computed by the method of token matching. The [N][N] matrix is formed where N is the total number of sentence in a document. The diagonal elements of a matrix are set to zero as the sentence should not be compared with itself. The similarity of each sentence pair is calculated as follows:

$$F6 = \frac{\sum[\mathrm{Sim}(Si, (Sj))]}{\mathrm{Max}[\mathrm{Sim}(Si, (Sj))]}$$

7. Term weight: The score of this feature is given by the ratio of summation of term frequencies of all terms in a sentence over the maximum of summation values of all sentences in a document. It is calculated by the following equation.

$$F7 = \frac{\sum TF_I}{MAX(\sum TF_I)}$$

Where, i=1 to n, n is the number of terms in a sentence.

8. Proper Nouns: The sentence that contains maximum number of proper nouns is considered to be important.

Its score is given by,

$$F8 = \frac{\text{Number of proper nouns in the sentence s}}{\text{Sentence length(s)}}$$

(1) $\alpha + \beta = \chi$. (1) (1)

C. Fuzzy Logic Scoring

Thus each sentence is associated with 8 feature vector. Using all the 8 feature scores, the score for each sentence are derived using fuzzy logic method. The fuzzy logic method uses the fuzzy rules and triangular membership function .The fuzzy rules are in the form of IF-THEN .The triangular membership function fuzzifies each score into one of 3 values that is LOW, MEDIUM & HIGH. Then we apply fuzzy rules to determine whether sentence is unimportant, average or important. This is also known as defuzzification.

For example IF (F1is H) and (F2 is M) and (F3 is H) and (F4 is M) and (F5 is M) and (F6 is M) and (F7 is H) and (F8 is H) THEN (sentence is important).

D. Sentence Selection

All the sentences of a document are ranked in a descending order based on their scores. Top n sentences of highest score are extracted as document summary based on compression rate. Finally the sentences in summary are arranged in the order they occur in the original document.

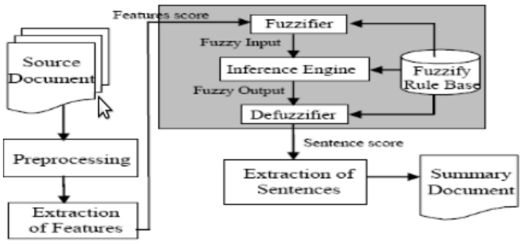

Fig. 1 Text Summarization using Fuzzy Inference System [6].

International Journal of Innovative Research in Advanced Engineering (IJIRAE) ISSN: 2349-2163
Volume 1 Issue 4 (May 2014) http://ijirae.com

V. LATENT SEMANTIC ANALYSIS

Latent Semantic Analysis (LSA) is a statistical model of word usage that permits comparisons of semantic similarity between pieces of textual information. LSA was originally designed to improve the effectiveness of information retrieval methods by performing retrieval based on the derived "semantic" content of words in a query as opposed to performing direct word matching. This approach avoids some of the problems of synonymy, in which different words can be used to describe the same semantic concept. The primary assumption of LSA is that there is some underlying or "latent" structure in the pattern of word usage across documents, and that statistical techniques can be used to estimate this latent structure. The term "documents" in this case, can be thought of as contexts in which words occur and also could be smaller text segments such as individual paragraphs or sentences. Through an analysis of the associations among words and documents, the method produces a representation in which words that are used in similar contexts will be more semantically associated.

All summarization methods based on LSA [18] use three main steps. These steps are as follows:

1. Input Matrix Creation.
2. Singular Value Decomposition.
3. Sentence Selection.

A. Input Matrix Creation

The input document is represented in a matrix form to perform the calculations.

A matrix which represents the input text is created. The columns of the matrix represent the sentences of the input text and the rows represent the words. The cells are filled out to represent the importance of words in sentences using different approaches, whose details are described in the rest of this section. The created matrix is sparse.

The first step of input matrix creation is to create the matrix in the form of terms x sentences. Assuming there are m terms and n sentences, the matrix A with size of m x n is created, which is A = [A1, A2, An]. Each column Ai represents weighted term vector of sentence i of the input document. The terms can be words/phrases that have been seen in the sentences, or they can be preprocessed before the creation of the matrix.

In order to reduce matrix size, the rows of the matrix the words can be reduced by preprocessing approaches like stop word removal, using roots of words only, using phrases instead of words and etc. These preprocessing approaches are mostly language dependent.

The cell is filled with the frequency of the word in the sentence.

B. Singular Value Decmposition

SVD is an algebraic method [18] that can model relationships among words/phrases and sentences.

Singular value decomposition is a mathematical method which models the relationships among terms and sentences. It decomposes the input matrix into three other matrices as follows:

$$A = U \sum V^T$$

A: Input matrix (m x n)
U: Words x Extracted Concepts (m x n)
$\sum$: Scaling values, diagonal descending matrix (n x n)
V^T: Sentences x Extracted Concepts (n x n)

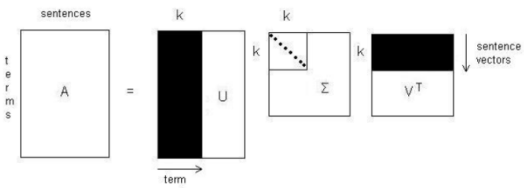

Fig. 2 Singular Value Decomposition [18]

C. Sentence Selection

Using the results of SVD, different algorithms use different approaches to select important sentences.
Different algorithms are proposed to select sentences from the input text for summarization using the results of SVD.
We are using Gong and Liu, 2001 summarization algorithm which use V^T matrix for sentence selection.
The algorithm of Gong and Liu is one of the main studies conducted in LSA based text summarization. After representing the input document in matrix and doing calculations of SVD, V^T matrix, matrix of extracted concepts x sentences, is used for selecting the important sentences. In V^T matrix, row order indicates the importance of the concepts such that the first row

represents the most important concept extracted. The cell values of this matrix show the relation between the sentence and the concept. A higher cell value indicates that the sentence is more related to the concept. In the approach of Gong and Liu [19], one sentence is chosen from the most important concept, and then second sentence is chosen from the second most important concept; and this process continues until all predefined number of sentences is collected. The number of sentences to be collected is given as a percentage of Summaries.

V^t matrix (r = 2)			
	Sent0	Sent1	Sent2
Con0	0,457	0,728	0,510
Con1	-0,770	0,037	0,637

Fig. 3 VT **matrix.**

From each row, sentence with highest score is chosen until predefined numbers of sentences are collected.

VI. PRAPOSED METHOD TO SUMMARIZATION

Traditional extraction methods cannot capture semantic relations between concepts in a text. Therefore, the use of semantic analysis to capture the semantic contents in sentences along with sentence extraction method brings into the improved summarization method. Our proposed method can improve the quality of summary with the help of latent semantic analysis and sentence feature extracted fuzzy logic system.

A. Proposed Architecture

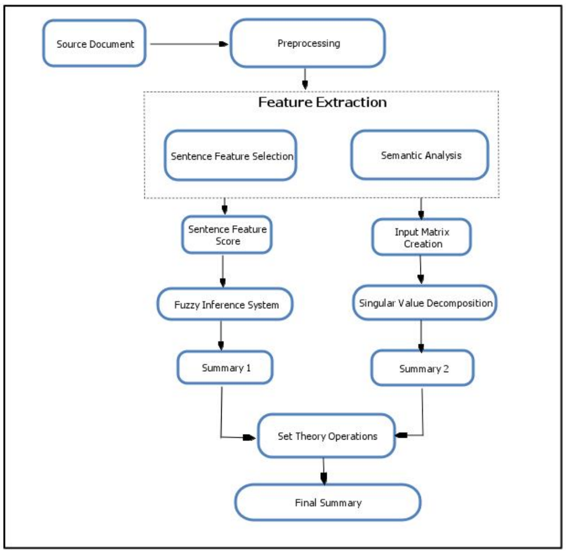

Fig. 4 Proposed Summary Architecture

The system consists of the following main steps:

(1) Read the source document into the system;
(2) For preprocessing step, the system extracts the individual sentences of the original documents. Then, separate the input document into individual words. Next, remove stop words. The last step for preprocessing is word stemming;
(3) In the sentence features extraction of fuzzy system each sentence is associated with vector of eight features that described in above Section, whose values are derived from the content of the sentence; In the same way to the semantic system the input matrix of term by documents is created with cell values.
(4) Then in the fuzzy system the features are calculated to obtain the sentence score by fuzzy logic method shows in Figure 1; and singular value decomposition is performed on input matrix to obtain the sentence by extracted concept matrix. i.e. importance of the concept & sentences related to that concept. Sentence with highest index value is the most representative sentences of that cpt.
(5) In fuzzy system, a set of highest score sentences are extracted as document summary based on the compression rate, and in SVD system, the V^T matrix cell values represents the most important sentences extracted . A higher cell value indicates the most related sentence .Thus numbers of sentences are collected into the summary based on compression rate.

(6) After getting summary1 and summary2, we intersect both summaries and extract a set of common sentences and a set of uncommon sentences. From uncommon set, we extract the sentences with high sentence scoring. And final set of improved summary is obtained by union of both the sets. We have incorporated the semantic contents of sentences into the sentence feature extraction Fuzzy summarization method to bring the much more improved summary.

B. Mathematical Expression

D = {s1 , s2 , s3 , s4,............s_n}
Where
- s is Sentences.

- D is Document.

If(W>T)
Where
- W is Weight.
- T is Threshold.

$Sum_{(old)} = \{s1 , s2\};$

$Sum_{(new)} = \{s1 , s3 , s4\};$

$Sum_{(Final)} = \{ Sum_{(old)} \wedge Sum_{(new)}\} \; U \; \{ W_P[Sum_{(old)} \; U \; Sum_{(new)}]\};$

Where .... (W>T)

$Sum_{(Final)} = \{s1\} \; U \; \{s4\} ;$

$\mathbf{Sum_{(Final)} = \{s1 , s4\};}$

Thus Summary is improved with our proposed method .Sentences are selected in the summary with the help of sentence scores. Higher scoring (ranked) sentences are added into the summary. Summary from Fuzzy system S1 and summary from LSA S2 are taken into account and common sentences are kept in one set and other sentences from S1 and S2 are chosen by their sentence scores. Sentences with high score are added into the summary.

VII. EVALUATION AND RESULTS

Old (Fuzzy-based) Summary	Proposed Summery(Fuzzy+LSA)	Online Summery
64.12	82.02	75.25
69.012	86.25	77.14
62.56	88.47	76.32
69.74	90.36	79.65
62.31	79.56	75.45

We use human generated summary as a gold summary standard for evaluation that has become standards of automatic evaluation of summaries. It compares the summaries generated by our method with the human generated (gold standard) summaries. For comparison, we used accuracy statistics. Our evaluation was done using accuracy percentage which was found to have the highest correlation with human judgments, namely, at a confidence level of 95%. It is claimed that our summary correlates highly with human assessments and has high recall and precision significance test with manual evaluation results. So we choose precision, recall as the measurement of our experiment results. In the table 1, we compare fuzzy summarizer (old Summary) with our proposed and online summarizer with different data sets.

Accuracy Table for Old Summary and Proposed Summary:

The results are shown in Table. The old summary with percentage of accuracy, the proposed summary with accuracy and online summary with the gold standard human generated summary.

The results are shown in Table 1; Old summary reaches an average accuracy of 65.5484, while online summarizer reaches the average of 76.762. Our proposed summarizer achieves the average accuracy of 85.332 with respect to the gold standard summary (human generated). Overall accuracy score from our summarizer are better than old (Fuzzy based) and online that shown in Figure 1.The another parameter to evaluate our summary is time complexity. As shown in Table 2, the time complexity increases as length of summary increases.

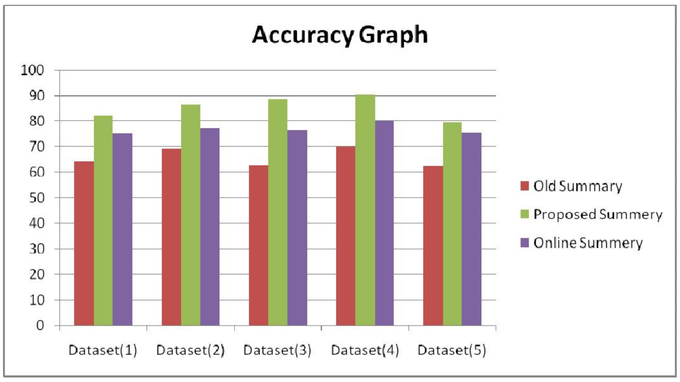

Fig. 5 Average accuracy result

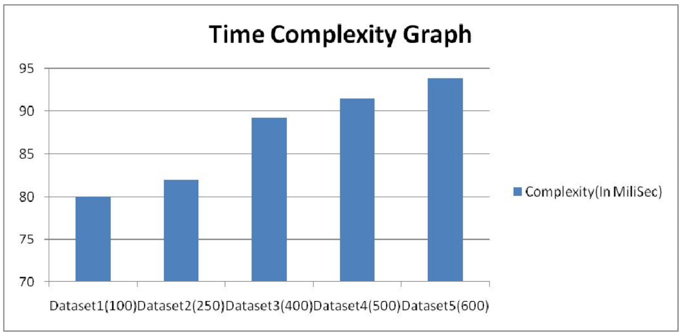

Fig. 6 Time Complexity result

The results are shown in Table. The old summary with percentage of accuracy, the proposed summary with accuracy and online summary with the gold standard human generated summary.

VIII. CONCLUSION

Automatic summarization is a complex task that consists of several sub-tasks. Each of the sub-tasks directly affects the ability to generate high quality summaries. In extraction based summarization the important part of the process is the identification of important relevant sentences of text. Use of fuzzy logic as a summarization sub-task improved the quality of summary by a great amount. The results are clearly visible in the comparison graphs. Our algorithm shows better results as compared to the output produced by two online summarizers.

Thus our proposed method improves the quality of summary by incorporating the latent semantic analysis into the sentence feature extracted fuzzy logic system to capture the semantic relations between concepts in the text. The focus of this paper is narrow: summarization of documents, but the ideas are more broadly applicable. We conclude that we need to extend the proposed method for multi document summarization with in a large data set.

REFERENCES

[1] Saeedeh Gholamrezazadeh ,Mohsen Amini Salehi, “A Comprehensive Survey on Text Summarization Systems ”, 978-1-4244-4946-0,2009 IEEE.

[2] Vishal Gupta and Gurpreet Singh Lehal "A survey of Text summarization techniques ",Journal of Emerging Technologies in Web Intelligence VOL 2 NO 3 August 2010.

[3] Oi Mean Foong ,Alan Oxley and Suziah Sulaiman "Challenges and Trends of Automatic Text Summarization ",International Journal of Information and Telecommunication Technology Vol.1, Issue 1, 2010.

[4] Archana AB, Sunitha. C ,"An Overview on Document Summarization Techniques" ,International Journal on Advanced Computer Theory and Engineering (IJACTE) ,ISSN (Print) : 2319 ˝U 2526, Volume-1, Issue-2, 2013 .

[5] Rafael Ferreira ,Luciano de Souza Cabral ,Rafael Dueire Lins ,Gabriel Pereira e Silva ,Fred Freitas ,George D.C. Cavalcanti ,Luciano Favaro , "Assessing sentence scoring techniques for extractive text summarization ",Expert Systems with Applications 40 (2013) 5755-5764 ,2013 Elsevier .

[6] L. Suanmali , N. Salim and M.S. Binwahlan,"Fuzzy Logic Based Method for Improving Text Summarization" , International Journal of Computer Science and Information Security, 2009, Vol. 2, No. 1,pp. 4-10.

[7] Mrs.A.R.Kulkarni , Dr.Mrs.S.S.Apte "A DOMAIN-SPECIFIC AUTOMATIC TEXT SUMMARIZATION USING FUZZY LOGIC ",International Journal of Computer Engineering and Technology (IJCET), ISSN 0976- 6367(Print), ISSN 0976 - 6375(Online) Volume 4, Issue 4, July-August (2013).

[8] Farshad Kyoomarsi ,Hamid Khosravi ,Esfandiar Eslami ,Pooya Khosravyan Dehkordy; "Optimizing Text Summarization Based on Fuzzy Logic ",Seventh IEEE/ACIS International Conference on Computer and Information Science ,978- 0-7695-3131-1 ,2008

[9] Ladda Suanmali ,Naomie Salim and Mohammed Salem Binwahla ,"Feature-Based Sentence Extraction Using Fuzzy Inference rules ",2009 International Conference on Signal Processing Systems ,978-0-7695-3654-5 ,2009 IEEE .

[10] Ladda Suanmali, Naomie Salim and Mohammed Salem Binwahlan "Fuzzy Genetic Semantic Based Text Summarization ", 2011 Ninth Ninth International Conference on Dependable, Autonomic and Secure Computing ,978-0-7695-4612-4 ,2011 IEEE .

[11] Ladda Suanmali, Mohammed Salem Binwahlan and Naomie Salim "Sentence Features Fusion for Text Summarization Using Fuzzy Logic ",2009 Ninth International Conference on Hybrid Intelligent Systems ,978-0-7695-3745-0 ,2009 IEEE.

[12] Hsun-Hui Huang ,Yau-Hwang Kuo ,Horng-Chang Yang ,"Fuzzy-Rough Set Aided Sentence Extraction Summarization",Proceedings of the First International Conference on Innovative Computing, Information and Control (ICICIC'06),0-7695- 2616-0/06 ,IEEE.

[13] Feifan Liu and Yang Liu, Member, IEEE "Exploring Correlation Between ROUGE and Human Evaluation on Meeting Summaries ",IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, VOL. 18, NO. 1, JANUARY 2010

[14] ZHANG Pei-ying ,LI Cun-he , "Automatic text summarization based on sentences clustering and extraction ",978-1-4244-4520-2 ,2009 IEEE .

[15] Udo Hahn ,Inderjeet Man ,"The Challenges of Automatic Summarization ",0018-9162/00,2000 IEEE .

[16] Róbert Móro, Mária Bieliková "Personalized Text Summarization Based on Important Terms Identification ",2012 23rd International Workshop on Database and Expert Sytems Applications ,1529-4188, 2012 IEEE .

[17] Rafael Ferreira ,Luciano de Souza Cabral ,Rafael Dueire Lins ,Gabriel Pereira e Silva , "Assessing sentence scoring techniques for extractive text summarization", Expert Systems with Applications 40 (2013) 5755-5764,,2013,Elsevier,Ltd.

[18] Josef Steinberger, Karel Ježek , "Using Latent Semantic Analysis in Text Summarization and Summary Evaluation", Department of Computer Science and Engineering, Univerzitní 22, CZ-306 14 Plzeň..

[19] MAKBULE GÜLÇİN ÖZSOY , " TEXT SUMMARIZATION USING LATENT SEMANTIC ANALYSIS ", A THESIS, THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES OF MIDDLE EASTTECHNICALUNIVERSITY.

