Multi-document English Text Summarization using Latent Semantic Analysis

Soniya Patil, Ashish T. Bhole

Abstract— In today's busy schedule, everybody expects to get the information in short but meaningful manner. Huge long documents consume more time to read. For this, we need summarization of document. Work has been done on Single-document but need of multiple document summarization is encouraging. Existing methods such as cluster approach, graph-based approach and fuzzy-based approach for multiple document summaries are improving. The statistical approach based on algebraic method is still topic of research. It demands for improvement in the approach by considering the limitations of Latent Semantic Analysis (LSA). Firstly, it reads only input text and does not consider world knowledge, for example women and lady it does not consider synonyms. Secondly, it does not consider word order, for example I will deliver to you tomorrow, deliver I will to you or tomorrow I will deliver to you. These different clauses may wrongly convey same meaning in different parts of document. Experimental results have overcomed the limitation and prove LSA with tf-idf method better in performance than KNN with tf-idf.

Index Terms—Natural language Processing (NLP), multi-document, Latent Semantic Analysis (LSA), Singular Value Decomposition (SVD).

1. INTRODUCTION

Natural Language Processing (NLP) is the computerized approach to examine text that is based on both a set of theories and a set of technologies.

Definition is defined as Natural Language Processing is a theoretically motivated range of computational techniques for examining and representing naturally occurring texts at one or more levels of language analysis for achieving language like Human for processing a range of tasks or applications.

The goal of automatic text summarization is to compress the given text to its necessary contents, based upon users' choice of shortness. In this system, the summary is generated to draw the most significant information in a shorter form of the source text, while still keeping its principal semantic content and helps the user to quickly understand large volumes of information.

Text Summarization methods can be classified into extractive and abstractive summarization. An extractive summarization method consists of picking important sentences, paragraphs etc. from the original document and concatenating them in brief. The importance of sentences is decided based on statistical and linguistic features of sentences. Abstractive summarizations try to develop an understanding of the main concepts in a document and then express those concepts in clear natural language. It uses linguistic methods to examine and read 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 from the original text document.

The single-document summarization task was approximately dropped. In multi-document summarization, important points are mixed up, such as reducing each document, combine all documents significant idea, compare the ideas from each, ordering sentences come from different sources keeping the logical and grammatical structure right[1]-[3].

Existing methods for Multi-Document summarization approaches like Graph-based, Fuzzy-based and Cluster-based are discovered. The algebraic approach consists of LSA method which is topic of research for multi-document text summarization. Its limitations are firstly that it reads only input text and does not consider world knowledge e.g. women and lady.

Secondly, it does not consider word order e. g. I will deliver to you tomorrow, deliver I will to you or tomorrow I will deliver to you. These clauses will be detected which wrongly convey same meaning.

Simply, multi-document text summarization means to retrieve salient information about a topic from various sources. Given a set of documents D = (D1, D2,…,Dn) on a topic T, the task of multi-document summarization is to identify a set of model units (S1,S2,..,Sn). The model units can be sentences, saying or generated semantically

- *Soniya Patil is Research Scholar in Department of Computer Engineering, S.S.B.T.'s College of Engineering & Technology, Jalgaon, Maharashtra, India. E-mail: patil.soniya2@gmail.com*
- *Ashish T. Bhole is working as Associate Professor in the Department of Computer Engineering at S.S.B.T.'s College of Engineering & Technology, Jalgaon, Maharashtra, India. E-mail: ashishbhole@hotmail.com*

correct language entity carrying some valuable information. Then important sentences are extracted from each model units and re-organized them to get multi-documents summary [3][4]. Summarization task can be classified into two types:

1) Single document text summarization.
2) Multi-document text summarization.

Process of multi-document summarization can be depicted in Figure 1.

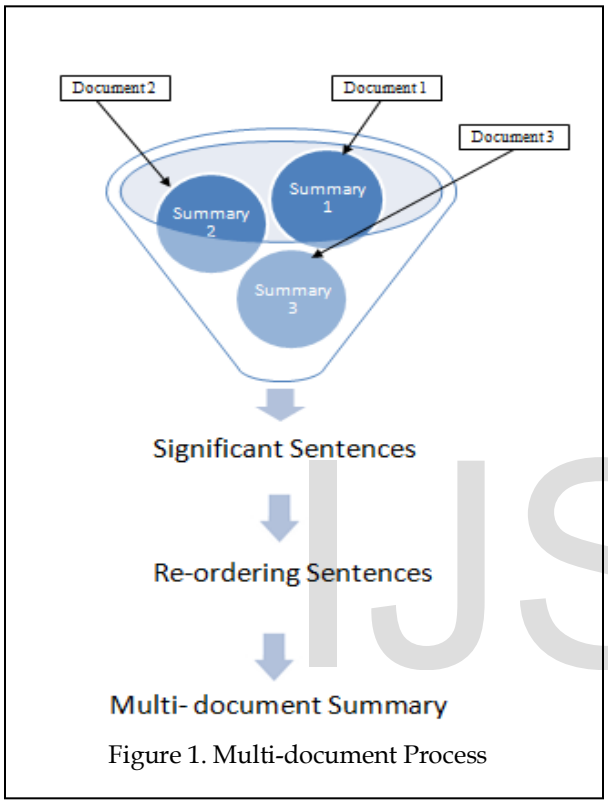

Figure 1. Multi-document Process

This paper is organized as follows: Section 1 introduces about Natural Language Processing area. Its motivation and problem definition. Related Work is described in Section 2. Section 3 introduces Proposed Work to overcome the limitation.

2. RELATED WORK

NLP began in the 1950s as the connection of artificial intelligence and linguistics. NLP was originally distinct from text information retrieval (IR), which employs highly scalable statistics-based techniques. Chomsky's 1956 theoretical analysis of language grammars provided an guess of the problem's difficulty, influence the creation (1963) of Backus- Naur Form (BNF) notation[5].

Thakkar and Chandak in 2010, compared two graph-based methods namely, Ranking algorithm and Shortest path algorithm. In this each sentence was assigned a node and accordingly same words were joined through edges. They concluded Shortest path algorithm was best suited as it generates smooth summaries in text form. Figure 2 shows the graph-based architecture [6].

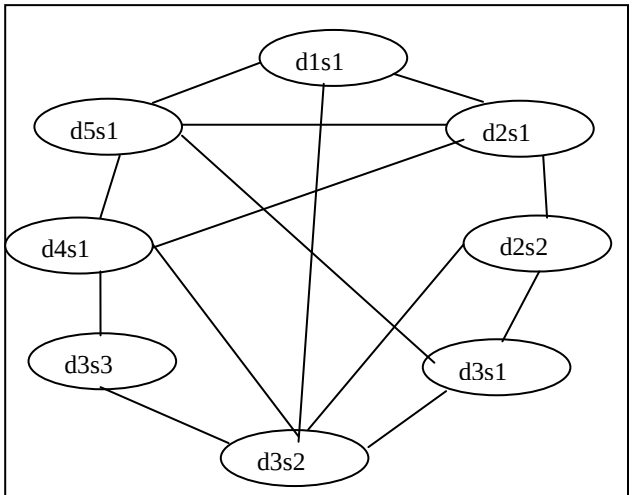

Figure 2. Graph-Based Architecture

But it sometimes it may happen most sentences come from same paragraph. Ozsoy, Cicekli in 2010, proposed LSA for multi-document for text summarization in Turkish language. In this LSA again explains its two approaches Cross and Topic which performs sentence selection. Its is used for Turkish language and sentence selection is done based on similarity of terms [7].

Chandra, Gupta and Paul in 2011, Statistical approach K-mixture Semantic Relationship Significance (KSRS). In this, similar terms are first weighted and then relationship is evaluated. Its summary extraction is 50% only[8].

Ladda, Salim and Mohammed in 2011, proposed Fuzzy Genetic approach shown in Figure 3, which is based on fuzzy IF-THEN rules and then applying fitness function on it[9]. The combination of methods is used, no single Fuzzy and Genetic could give such output for Multi- document. Due to combination of methods complexity increased.

Nguyen, Pham and Doan in 2012, proposed Genetic Programming that ranks the sentences based on their importance and applies fitness function on it. It's not suitable for English Documents[10].

Asef, Kahani, Yazdi and Kamyar in 2011, proposed LSA for multi-document for text summarization for Persian language. In this LSA again it performs term selection. Its limitation: used for Persian language differs from English language both morphologically and semantically [11].

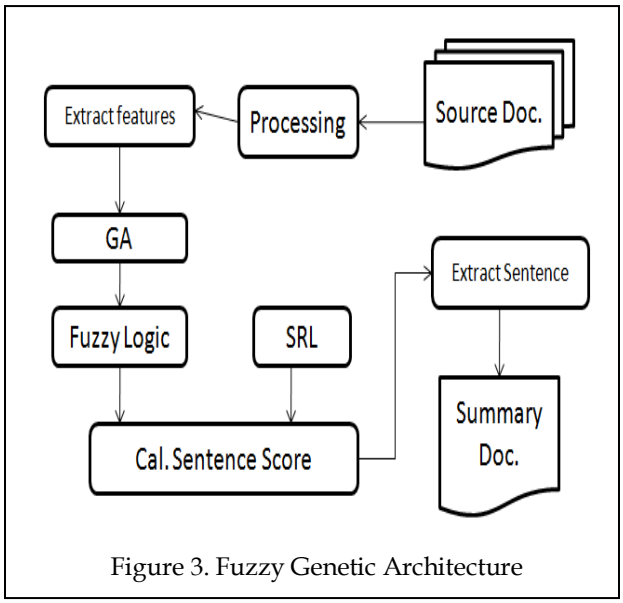

Figure 3. Fuzzy Genetic Architecture

Xuan Li, Liang Du, and Yi-Dong Shen in May 2013, proposed an improved model of Graph-based on Ranking algorithm. It considers Group of sentences. Limitation: its NP-hard method so approximation takes place [12]. Table 1 shows the detail literature survey.

TABLE 1
LITERATURE SURVEY

Sr. No.	Category	Author and Year	Proposed Concept	Limitation
1	Graph Based Approach	Thakkar and Chandak, 2010	Each sentence was assigned a node and same words were joined through edges	Most sentences occur from same paragraph
		Xuan Li, Liang Du, Yi-Dong Shen, 2013	Ranking algorithm	NP-hard method
2	Fuzzy and Genetic Approach	Ladda, Samir, Mohammed, 2011	IF-THEN rules and Fitness function	Complexity increased
3	Genetic Approach	Nyugen, Pham and Doan, 2012	Ranks sentences and fitness function	Not suitable for English documents
4	Statistical approach	Ozsoy, Cicekli,2010	LSA for Turkish language	Sentences are selected based on similarity
		Chandra, Gupta and Paul, 2011	K-mixture semantic Relationship significance	Output is 50% only
		Asaf, Kahani, Yazdi and Kamyar, 2011	LSA for Persian language	Persian language differs from English lang. both morphologically and semantically

3. PROPOSED WORK

This section introduces about the new advancement in LSA method for improving its limitations.

The design in figure 4 describes about execution. What should be its input material, how does it process and what is its desired output. Multiple documents are given as input. Then, it extracts the sentences based on term frequency taking into account its meaning and distributes as words among abstract and from other sections. The first step in the process it to form the numerical dataset by collecting documents.

LSA is based on the Vector Space Model, every document is signified by a vector in a highly dimensional space and every element in the vector stand for the weight for a given term for the document at hand.

A. Pre-processing/ Training phase: Text preprocessing step is an important step which trains the system for identification and making directory.

The first step in pre-processing phase is tokenization. Tokenization is the process of breaking a stream of text up into words, phrases, symbols, or other meaningful elements called tokens. The list of tokens becomes input for further processing. Tokenization is useful both in linguistics (where it is a form of text segmentation), and in computer science. In the second step, All common words that do not add to the individual meaning and situation of documents can be removed before indexing (e.g. "a", "the"). Universally a used word lists are available including a large set of so-called 'stop' words Stop words are being removed from the document. Some elements like articles; short verbs etc which are considered as a stop word are listed in a file to be eliminated. In next step, the idea of stemming is to improve the ability to detect similarity not considering the use of word alternative (stemming reduces the number of synonyms, since multiple terms sharing the same stem are mapped onto the same concept or stem). In the next step, after removing redundancy of words, dictionary is prepared and tf-idf matrix is formed [13][14].

B. Sentence Selection:

1. Extracting the existing concept of documents: In this phase, LSA has been used for extracting the main concepts of the document. Then, Singular Vector Decomposition (SVD) is used as a rank lowering method to truncate the original vector.

It will decompose the original term-by-document matrix into orthogonal factors that represent both terms and documents.SVD function performs matrix vectorization.

2. The cosine distance between the concept vector and the document vector is calculated. This value represents the amount of similarity of each concept with a topic in the framework. In the other words, main concept of the topic is extracted. Minimum the cosine similarity value, nearest or identical is that document to the test file.

3. Selecting file with highest index value of keyword: In this step, calculate the frequency of keywords in each file document and depending on the greatest value of particular file, summary is being created.

The mathematical formulae used in this implementation are shown below. The very important function of SVD is used which helps to identify the document which belong to particular test file.

The distinctive feature of SVD is that it is capable of capturing and representing interrelationships among terms so that they can semantically cluster terms and sentences.

$$A = U\sum V^T \quad (1)$$

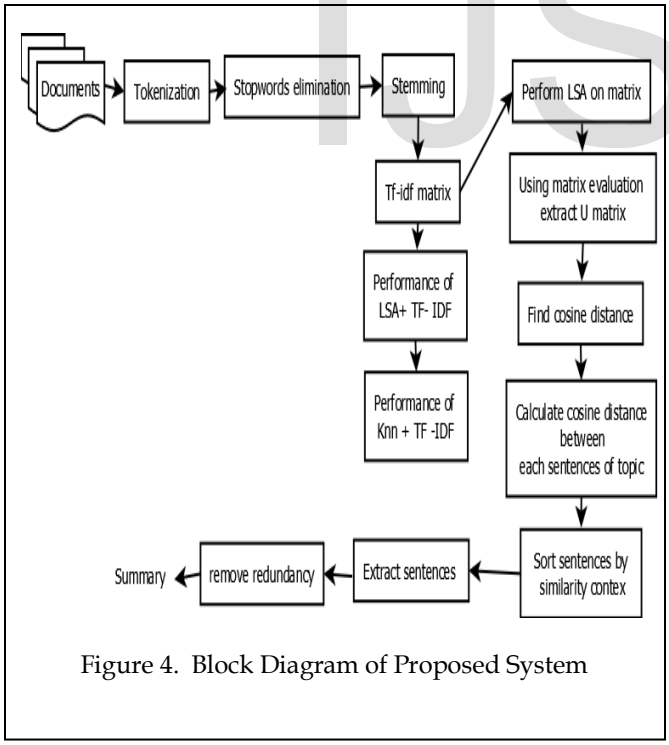

Figure 4. Block Diagram of Proposed System

Term Frequency- Inverse Document Frequency method determine the relative frequency of words in a specific document through an inverse proportion of the word over the entire document quantity [13].

$$(tf - idf)_{ij} = tf_{ij} * \log_2(\frac{N}{df_i}) \quad (2)$$

Cosine distance is calculated between column vectors of matrix U (Ci) and document vector (Dj).

$$\cos(Ci, Dj) = \frac{\sum_k C_{ik} d_{jk}}{\sqrt{\sum_k (C_{ik})^2} \times \sqrt{\sum_k (d_{jk})^2}} \quad (3)$$

4. RESULTS

The Experimental results of the simulation shows the following observations between the proposed LSA with tf-idf and existing tf-idf, difference between proposed method and KNN with tf-df and lastly difference between proposed method and copernic summarizer tool.

Contingency table is denoted which represents the relation about the documents for calculating recall, precision and accuracy of method.

1) TPi (True Positive): number of correctly classified documents as in Ci , which belong to the class Ci.

2) FPi (False Positive): number of incorrectly classified documents as in Ci , which do not belong to the class Ci.

3) FNi (False Negative): number of incorrectly classified documents as in not Ci, which are in the class Ci.

4)TNi (True Negative): number of correctly classified documents as in not Ci, which are not in the class Ci[14].

TABLE 2
CONTIGENCY TABLE

	Belong to C_i	Not belong to Ci
Classified to class Ci	TP_i	FP_i
Not classified to Ci	FN_i	TN_i

Recall Index refers to how many documents truly belonging to same category have been classified in class Ci.

$$\text{Recall}_{Ci} = \frac{TPi}{TPi + FNi}$$

Precision is the ratio of documents classified correctly in the class Ci with the documents assigned to the class Ci.

$$\text{Precision}_{Ci} = \frac{TPi}{TPi + FPi}$$

Accuracy: it refers to the ratio of documents classified correctly to the class Ci and other than Ci among all the documents.

$$\text{Accuracy}_{Ci} = \frac{TPi + TNi}{TPi + FPi + TNi + FNi}$$

TABLE 3: RECALL, PRECISION AND ACCURACY CLASSIFICATION INDEXES OF PROPOSED LSA WITH TF-IDF

Class	Recall Index	Precision Index	Accuracy Index
Currency	0.111	0.058	0.905
Military	0.05	0.166	0.897
Microsoft	0.08	1.000	0.909
News story	0.04	0.2	0.905
USA	0.0416	0.5	0.905
Politics	0.04	1.000	0.905
Politics	0.0416	0.5	0.905
Entertainment	0.142	0.28	0.929
International	0.0416	0.333	0.905
International	0.0416	0.333	0.905

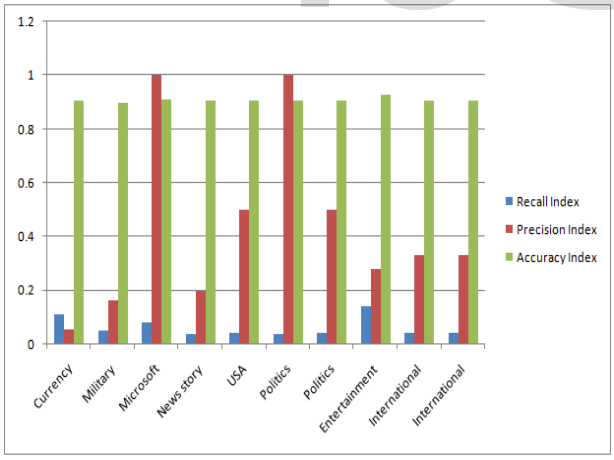

Figure 5: Classification indexes of different dataset

TABLE 4: COMPARISON BETWEEN METHODS OF RECALL INDEX

	Methods	
	TF-IDF	Proposed LSA with TF-IDF
Recall Index (20% summary proportion)	0.5-0.6	0.06

TABLE 5: COMPARISON BETWEEN METHODS OF PRECISION INDEX

	Methods	
	TF-IDF	Proposed LSA with TF-IDF
Precision Index (20% summary proportion)	0.5-0.6	0.4

TABLE 6: COMPARISON BETWEEN METHODS OF PRECISION INDEX

	Methods	
	TF-IDF	Proposed LSA with TF-IDF
Accuracy Index (20% summary proportion)	0.5-0.6	0.907

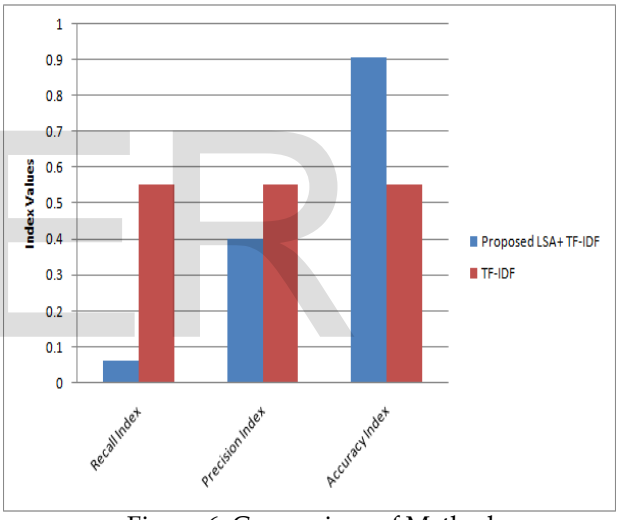

Figure 6: Comparison of Methods

TABLE 7: COMPARISON BETWEEN LSA WITH TF-IDF AND KNN WITH TF-IDF

Parameters	LSA +Tf-idf	KNN + Tf-idf
Input documents (% of identical)	100%	20%
Run time in seconds	172	220

Table 7 and Table 8 shows the comparison of other methods and tool with proposed system.

5. CONCLUSION

In this paper LSA method has been combined with tf-idf , in which SVD plays role of matrix decomposition. Tf-idf has helped to calculate and make word dictionary for forming keyterms. The keyterms selection is proved to be advantageous over existing sentence selection.

The accuracy of existing system is 50% - 60% whereas proposed system has 90%. The KNN with tf-idf method identifies only 20% of input document whereas proposed method identifies 100% input document. Copernic summarizer tool requires input file in pdf format only whereas proposed system takes input in directory containing text files.

TABLE 8: COMPARISON BETWEEN PROPOSED SYSTEM AND COPERNIC SUMMARIZER TOOL

Parameters	Proposed System	Copernic summarizer Tool
Type of Input document	Directory with PDF files	required text files only
Form of output	Paragraph	Line by line document sentences
Requirement of summary proportion	not required	required as % of summary proportion
Number of words in summary	Between 20 to 30 words	More than 30 words

REFERENCES

[1] R. M. Badry, A. S. Eldin, and D. S. Elzanfally, "Text summarization within the latent semantic analysis framework: Comparative study," International Journal of Computer Applications (0975 8887), vol. 81, no. 11, pp. 40-43, November 2013.

[2] Elena Lloret, "Text Summarization : An Overview",pp 1- 24.

[3] Josef Steinberger, Karel Jezek, "Evaluation Measures For Text Summarization",Computing and Informatics,Vol. 28, 2009, 1001-1026, V 2009-Mar-2, pp.1002-1025.

[4] Josef Steinberger, Karel Jezek, "Using Latent Semantic Analysis in Text Summarization and Summary Evaluation".

[5] Prakash M Nadkarni, Lucila Ohno-Machado, Wendy W Chapman, "Natural language processing: an introduction, J Am Med Inform Assoc 2011;18:544e551.

[6] Vishal Gupta, Gurpreet Singh Lehal, "A Survey of Text Summarization Extractive Techniques",Journal Of Emerging Technologies In Web Intelligence, Vol.2, no. 3, August 2010,Pp.258-268.

[7] Thakkar, Chandak,Dharaskar, "Graph-Based Algorithms for Text Summarization",IEEE Third International Conference,2009,pp.516-519.

[8] Ozsoy, Cicekli, "Text Summarization of Turkish Texts using Latent Semantic Analysis",23rd International Conference,Beijing,2010,pp.869876.

[9] Chandra, Gupta and Paul, "A Statistical approach for Automatic Text Summarization by Extraction",IEEE International Conference,2011,pp.268-271.

[10] Ladda, Salim and Mohammed, "Fuzzy Genetic Semantic Based Text Summarization",IEEE Ninth International Conference,2011, pp.1184-1190.

[11] Nguyen, Pham and Doan, "A Study on Use of Genetic Programming for Automatic Text Summarization",IEEE Fourth International Conference,2012, pp.93-97.

[12] Asef, Kahani, Yazdi and Kamyar,\Context-Based Persian Multi-document Summarization",IEEE International Conference,2011, pp.145-149.

[13] Xuan Li, Liang Du, and Yi-Dong Shen, "Update Summarization via Graph-Based Sentence Ranking",IEEE Transactions on Knowledge and Data Engineering,vol. 25, no. 5, may 2013,pp.1162-1174.

[14] Bruno Trstenjaka,Sasa Mikac, Dzenana Donko, "KNN with TF-IDF Based Framework for Text categorization",24th DAAAM International Symposium on Intelligent manufacturing and Automation,2013,Procedia Engineering 69(2014),pp 1356 1364.

[15] Naohiro Ishii, Tsuyoshi Murai,Takahiro Yamada,Yongguang Bao, "Text Classi_cation by Combining Grouping, LSA and kNN", 5th IEEE/ACIS International Conference on Computer and Information Science.