



Review of Multi-document Text Summarization Methods

¹Soniya Patil, ²Ashish T. Bhole
¹Research Scholar, ²Associate Professor
Department of Computer Engineering
SSBT's College of Engineering & Technology
Jalgaon, Maharashtra, INDIA

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 like graph-based approach, fuzzy-genetic based approach, genetic approach and statistical approach for multiple document summaries are improving. The statistical approach based on algebraic method is still the topic of research. Effort has to be done to improve this approach by considering the limitations of Latent Semantic Analysis (LSA). Firstly, it reads only input text and does not consider world knowledge. Secondly, it does not consider word order. The different clauses may convey same meaning in different parts of document. Lastly, all the approaches give their output in text form. The approach is to bring the output in tabular form, so that LSA method can be improved.

Keywords: Natural language Processing (NLP); Text Summarization; Latent Semantic Analysis (LSA)

I. 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. Natural Language Processing is defined as a theoretically motivated range of computational techniques for analyzing and representing naturally occurring texts at one or more levels of linguistic analysis for the purpose of achieving human-like language processing for a range of tasks or applications. The goal of automatic text summarization is to compress the given text to its necessary contents, based upon user's choice of briefness. In this system, the summary is generated by computer 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.

A summary can be employed in an indicative way as a pointer to some parts of the original document, or in an informative way to cover all relevant information of the text. In both cases the most important advantage of using a summary is its reduced reading time. Text Summarization methods can be classified into extractive and abstractive summarization.

An extractive summarization method consists of selecting important sentences, paragraphs etc. from the original document and concatenating them into shorter form. The importance of sentences is decided based on statistical and linguistic features of sentences. An Abstractive summarization attempts 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 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 from the original text document.

The single-document summarization task was approximately dropped. In multi-document summarization, several key points are involved, such as reducing each document, integrate all documents significant idea, compare the ideas found from each, ordering sentences come from different sources keeping the logical and grammatical structure right^[1].

Existing methods for Multi-Document summarization approaches like Graph-based, Fuzzy based, cluster based are discovered. But Algebraic approach consists of LSA method which is topic of research for multi-document. Its limitations are firstly that it reads only input text and does not consider world knowledge for example women and lady. Secondly, does not consider word order. For example, I will deliver to you tomorrow. deliver i will to you. Tomorrow i will deliver to you.

Simply, multi-document text summarization means to retrieve salient information about a topic from various sources. Given a set of documents $D = (D_1, D_2, \dots, D_n)$ on a topic T , the task of multi-document summarization is to identify a set of model units $S = (S_1, S_2, \dots, S_n)$. The model units can be sentences, phrases or some generated semantically correct language units carrying some useful information. Then significant sentences are extracted

from each model units and re-organized them to get multi-documents summary. 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 Figure1.

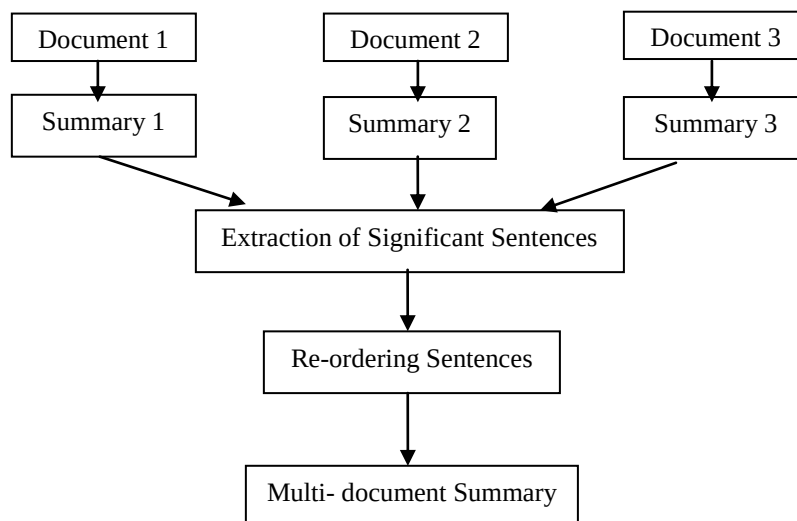


Figure 1: Multi-Document Process

Section I introduces the concept of Natural Language Processing. The motivation, problem definition and related work is described in Section II. Section III specifies the proposed work to overcome the limitation.

II. 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 a guess of the problem's difficulty, influence the creation (1963) of Backus- Naur Form (BNF) notation.

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 as shown in Figure 2^[2]. Its Limitation was that 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 limitation is that it is used for Turkish language and sentence selection is done based on similarity of terms^[3].

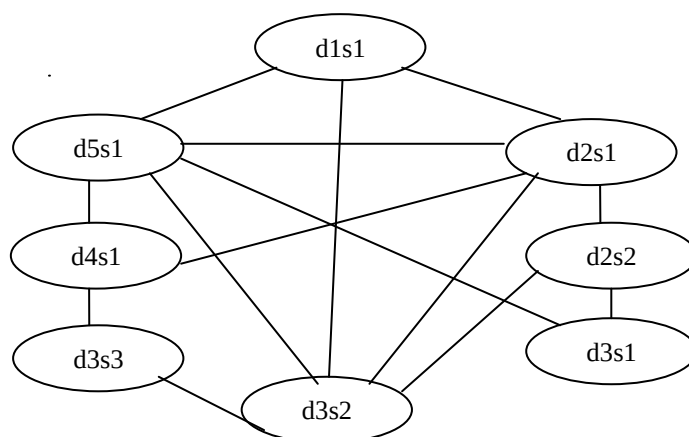


Figure 2: Graph-Based Architecture

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 limitation is summary extraction is 50% only^[4].

Ladda, Salim and Mohammed in 2011, proposed Fuzzy Genetic approach which is based on fuzzy IF-THEN rules as shown in Figure 3 and then applying fitness function on it^[5]. Its Limitation observed is that 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 is not suitable for English Documents^[6].

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^[7].

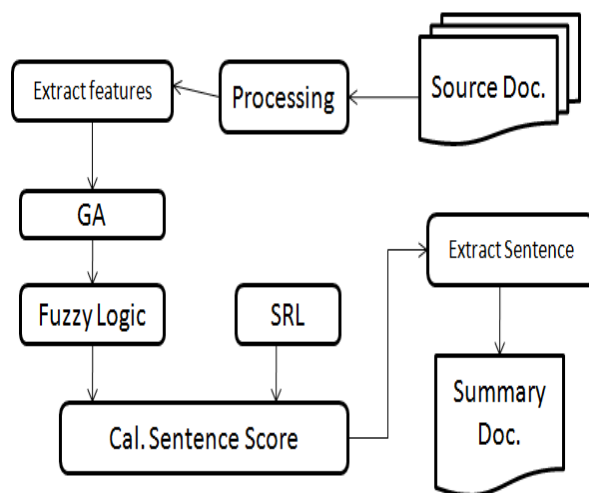


Figure 3: Fuzzy Genetic Architecture

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^[8].

The comparison of text summarization methods of different approaches is given in Table I.

Table I: Comparing Text Summarization Methods

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

III. Proposed Work

This section introduces the advancement in LSA method for improving its limitations. LSA is a fully automatic mathematical/statistical technique for extracting and inferring relations of expected contextual usage of words in passages of discourse. It is not a traditional natural language processing or artificial intelligence program; it uses no humanly constructed dictionaries, knowledge bases, semantic networks, grammars, syntactic parsers, or morphologies, or the like, and takes as its input only raw text parsed into words defined as unique character strings and separated into meaningful passages or samples such as sentences or paragraphs.

LSA existing method reads only input text and does not consider the world knowledge. It is equally important to consider the world knowledge because synonyms may be used in different other paragraphs. This limitation could be overcome by considering World Approved Dictionary. Secondly, LSA does not consider word order. Phrases containing same words but in different order used in different way were not considered, this limitation would be overcome by considering parser tree. Thirdly, all the outputs are in form of text summarization. The attempt is made to increase the format of output by converting text input into table or graph as output.

Block Diagram for working of LSA method Text summarization is shown in Figure 4. Text documents will be given as input. Text extraction will take place in process and accordingly sentences will be extracted from abstract and other part of documents using SVD (Singular Value Decomposition) method, semantic and syntactic analysis will be done and output will be gained.

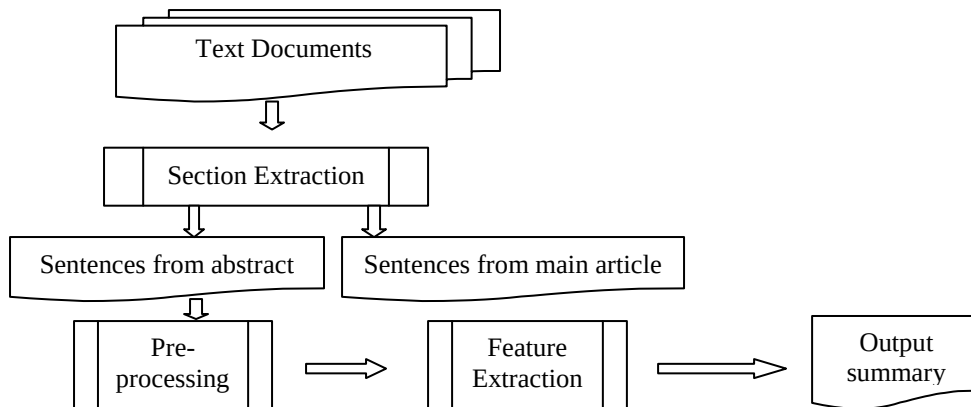


Figure 4. Block Diagram of Proposed Method for Multi- document Text Summarization

To construct representations of text sentences in the topic space, LSA uses matrix factorizations, such as singular value decomposition (SVD). The SVD of matrix A is defined as

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

Where,

$U = [u_{ij}]$ is an $m \times n$ column-orthonormal matrix whose columns are called left singular vectors;

$\Sigma = \text{diag}(\sigma_1, \sigma_2, \dots, \sigma_n)$ is an $n \times n$ diagonal matrix, whose diagonal elements are non-negative singular values sorted in descending order and

$V = [v_{ij}]$ is an $n \times n$ orthonormal matrix, whose columns are called right singular vectors.

Latent semantic analysis works with a collection of all text sentences. The original text is represented in the form of a numerical matrix as shown in Equation 1. Matrix columns correspond to text sentences and each sentence is represented as a vector a of fixed dimension n , where n is the number of text features and the i_{th} element of the vector determines weight of the i_{th} feature. To define a representation model, it is required to define a feature space and select a method of weight calculation.

$$\begin{bmatrix} a_{1,1} & a_{1,2} & a_{1,n} \\ a_{2,1} & a_{2,2} & a_{2,n} \\ \dots & \dots & \dots \\ a_{m,1} & a_{m,2} & a_{m,n} \end{bmatrix} = \begin{bmatrix} u_{1,1} & u_{1,k} \\ u_{2,1} & u_{2,k} \\ \dots & \dots \\ u_{m,1} & u_{m,k} \end{bmatrix} \begin{bmatrix} \sigma_1 & \dots & 0 \\ \dots & \sigma_k & \dots \end{bmatrix} \begin{bmatrix} v_{1,1} & v_{n,1} \\ v_{1,k} & v_{n,k} \end{bmatrix} \quad \text{---(1)}$$

IV. Conclusion

The improvement in LSA method will surely upgrade its importance. Now-a-days everyone tries to meet perfect blend of summarization as outcome after reading multiple documents. The limitations of existing statistical approach can be overcome and better output can be achieved using LSA method.

References

- [1] Md. Majharul Haque, Suraiya Pervin and Zerina Begum, "Literature Review of Automatic Multiple Documents Text Summarization", *International Journal of Innovation and Applied Studies*, Vol. 3, No. 1, pp. 121-129, May 2013.
- [2] Thakkar, Chandak, Dharaskar, "Graph-Based Algorithms for Text Summarization", *In 3rd International Conference on Emerging Trends in Engineering and Technology (ICETET)*, pp. 516-519, 2010.
- [3] Ozsoy, Cicekli, "Text Summarization of Turkish Texts using Latent Semantic Analysis", *In Proceedings of the 23rd International Conference on Computational Linguistics*, Beijing, pp. 869-876, 2010.
- [4] Chandra, Gupta and Paul, "A Statistical approach for Automatic Text Summarization by Extraction", *in Proceedings of International Conference on Communication Systems and Network Technologies (CSNT)*, pp. 268-271, 2011.
- [5] Ladda, Salim and Mohammed, "Fuzzy Genetic Semantic Based Text Summarization", *In Ninth IEEE International Conference on Dependable, Autonomic and Secure Computing (DASC)*, pp. 1184-1190, 2011.
- [6] Nguyen, Pham and Doan, "A Study on Use of Genetic Programming for Automatic Text Summarization", *in Fourth International Conference on Knowledge and Systems Engineering*, pp. 93-97, 2012.
- [7] Asef, Kahani, Yazdi and Kamyar, "Context-Based Persian Multi-document Summarization", *In International Conference on Asian Language Processing (IALP)*, pp. 145-149, 2011.
- [8] Xuan Li, Liang Du and Yi-Dong Shen, "Update Summarization via Graph-Based Sentence Ranking", *in IEEE Transactions On Knowledge and Data Engineering*, Vol. 25, No. 5, pp. 1162-1174, May 2013.