IJCSE International Journal of Computer Sciences and Engineering Open Access
Review Paper Volume-3, Issue-7 E-ISSN: 2347-2693

A Review on Automatic Text Summarization Techniques in NLP

Kirtipreet kaur[1]* and Deepinderjeet Kaur[2]

[1*,2] *Department of Computer Science, Desh Bhagat University Punjab, INDIA*

www.ijcseonline.org

Received: Jun/16/2015 Revised: Jun/22/2015 Accepted: July/15/2015 Published: July/30/ 2015

Abstract— In this article we present a survey on different text summarization techniques in natural language processing. Text Summarization is condensing the source text into a shorter version preserving its information content and overall meaning. It is very difficult for human beings to manually summarize large documents of text. There are many techniques of doing text summarization i.e. some are extractive as well as abstractive techniques. But we need that technique which will give meaningful summary without showing any redundancy or any type of ambiguity whether the summary will contain original text or not.

Keywords— Text Summarization, WordNet, Abstractive, Extractive.

I. INTRODUCTION

NLP is the branch of computer science focused on developing systems that allow computers to communicate with people using everyday language. NLP automates the translation process between computers and humans. It is a method of getting a computer to understandably read a line of text without the computer being fed some sort of clue or calculation. The foundations of NLP lie in a number of disciplines, viz. computer and information sciences, linguistics, mathematics, electrical and electronic engineering, artificial intelligence and robotics, psychology, etc. Applications of NLP include a number of fields of studies, such as machine translation, natural language text processing and summarization, user interfaces, multilingual and cross language information retrieval (CLIR), speech recognition, artificial intelligence and expert systems, and so on.

Automatic text summarization is the technique, where a computer summarizes a text. A text is entered into the computer and a summarized text is returned, which is a non redundant extract from the original text. These days, the number of Web pages on the Internet almost doubles every year as the information is now available from a variety of sources. It takes considerable amount of time to find the relevant information. Automatic Text Summarization will help the users to find the relevant information rapidly. An example of the use of summarization technology is search engines such as Google.

One of the natural questions to ask in summarization is "What are the texts that should be represented or kept in a summary?" The summary must be generated by selecting the important contents or conclusions in the original text. Finding out important information becomes a truly challenging task. Currently, the need for automatic text summarization has appeared in many areas such as news articles summary, email summary, short message news on mobile, and information summary for businessman, government officials, research, and online search engines to receive the summary of pages found and so on.

The most popular categorization of summary into single document and multi-document summarization. Single document is the process of creating a summary from a single text document. Multi-document summarization shortens a collection of related documents; into single summary.

Generally, there are two approaches to automatic summarization

- Extractive methods work by selecting a subset of existing words, phrases, or sentences in the original text to form the summary.
- Abstractive methods build an internal semantic representation and then use natural language generation techniques to create a summary that is closer to what a human might generate. Such a summary might contain words not explicitly present in the original.

II. LITERATURE SURVEY

E.PadmaLahari, D.V.N.Siva Kumar, S. Shiva Prasad [1] said that the text summarization is an emerging technique for finding out the summary of the text document. In this paper, they propose an automatic text summarization technique using both linguistic and statistical features using successive threshold for finding the summary i.e. important sentences from the given input text document. Here the sentences are selected for summary based on the weight of the sentence. The weight

of the sentences is calculated based on the statistical and linguistic features. Their approach assigns scores to the sentences by weighting the features like term frequency, word occurrences, and noun weight, phrases etc. In this approach, the number of sentences present in our summary would be equal to the number of paragraphs present in a text document, which can be achieved by using our successive threshold approach.

Alok Ranjan Pal, Diganta Saha [2] said that the text Summarization is the procedure by which the significant portions of a text are retrieved. Most of the approaches perform the summarization based on some hand tagged rules, such as format of the writing of a sentence, position of a sentence in the text, frequency of few particular words in a sentence etc. But according to different input sources, these predefined constraints greatly affect the result. The proposed approach performs the summarization task by unsupervised learning methodology. The importance of a sentence in an input text is evaluated by the help of Simplified Lesk algorithm. As an online semantic dictionary WordNet is used. First, this approach evaluates the weights of all the sentences of a text separately using the Simplified Lesk algorithm and arranges them in decreasing order according to their weights. Next, according to the given percentage of summarization, a particular number of sentences are selected from that ordered list. The proposed approach gives best results up to 50% summarization of the original text and gives satisfactory result even up to 25% summarization of the original text.

Ms.Pallavi D.Patil, Prof.N.J.Kulkarni [3] presented an algorithm using fuzzy logic. 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. This paper focuses on the Fuzzy logic Extraction approach for text summarization.

D.Y. Sakhare, Dr. Raj Kumar [4] said that recently, there has been a significant research in automatic text summarization using feature-based techniques in which most of them utilized any one of the soft computing techniques. But, making use of syntactic structure of the sentences for text summarization has not widely applied due to its difficulty of handling it in summarization process. On the other hand, feature-based technique available in the literature showed efficient results in most of the techniques. So, combining syntactic structure into the feature-based techniques is surely smooth the summarization process in a way that the efficiency can be achieved.

A.R.Kulkarni, S.S.Apte [5] presented a better algorithm such that this paper proposes a better approach for text summarization using lexical chaining and correlation of sentences. Lexical chains are created using Wordnet. The score of each Lexical chain is calculated based on keyword strength, Tf-idf & other features. The concept of using lexical chains helps to analyze the document semantically and the concept of correlation of sentences helps to consider the relation of sentence with preceding or succeeding sentence. In this paper they discuss a summarization method, which combines lexical chaining with correlation of sentences in which relation of a sentence with the preceding sentence is considered.

Atif Khan, Naomie Salim [6] presented a paper that it is very difficult for human beings to manually summarize large documents of text**.** Automatic abstractive summarization provides the required solution but it is a challenging task because it requires deeper analysis of text. In this paper, a survey on abstractive text summarization methods has been presented. Abstractive summarization methods are classified into two categories i.e. structured based approach and semantic based approach. The main idea behind these methods has been discussed. Besides the main idea, the strengths and weaknesses of each method have also been highlighted. Some open research issues in abstractive summarization have been identified and will address for future research. Finally, it is concluded from the literature studies that most of the abstractive summarization methods produces highly coherent, cohesive, information rich and less redundant summary.

Anjal R.Deshpande, Lobo L. M. R. J [8] presented a paper according to them; a summarization system consists of reduction of a text document to generate a new form which conveys the key meaning of the contained text. Due to the problem of information overload, access to sound and correctly-developed summaries is necessary. Text summarization is the most challenging task in information retrieval. Data reduction helps a user to find required information quickly without wasting time and effort in reading the whole document collection. This paper presents a combined approach to document and sentence clustering as an extractive technique of summarization.

Mohsen Pourvali and Mohammad Saniee Abadeh [9] said that the technology of automatic document summarization is maturing and may provide a solution to the information overload problem. Nowadays, document summarization plays an important role in information retrieval. In this paper we consider the effect of the use of lexical cohesion features in Summarization, And presenting a algorithm base on the knowledge base. Ours algorithm at first find the correct sense of any word, Then constructs the lexical chains, remove Lexical chains that less score than other ,detects topics roughly from lexical chains, segments the text with respect to the topics and selects the most important sentences.

Manisha Prabhakar, Nidhi Chandra [10] presented a paper based on pragmatic analysis .In this paper, text

summarization technique is designed for the documents having the fixed format. The proposed system generates the summary of the fixed format documents by analyzing all the different parts of the documents. The system consists of five stages. In first stage each sentence is partitioned into the list of tokens and stop words are removed. In second stage, frequency usage is counted for each word. In third stage, assign POS tag for each weighted term and Word sense disambiguation is done. In the fourth stage, pragmatic analysis is performed. After Pragmatic Analysis, summarized sentences will be store in a database.

III. PROPOSED WORK

The proposed system will work on single document summarization. The system will contain dictionary having meaningful keywords with every word to get meaningful sentences of the text to be summarized and we will use PHP language for this. The system will also generate a summary in English as well as in Hindi Language because there is very less work of text summarization is done in Hindi language.

IV. CONCLUSION

We have discussed an overview of the existing text summarization techniques in NLP. We have found that text summarization is still a raw area of research as seen many problems found in their produced summaries. Finally, it is concluded from the literature studies that most of the extractive summarization methods produce summaries easily but produce redundant summaries and many of them use wordnet which is a lexical database for giving sense of a word but still don't give accurate results as it contains limited information and abstractive summarization methods produces highly coherent, cohesive, information rich and less redundant summary but difficult to produce as they need NLG which itself is a growing field. Since the selection of the right procedure of removing problems occurred in both techniques plays an important role, it is important to experiment and a hybrid technique need to be developed which uses both techniques to get meaningful summary.

ACKNOWLEDGEMENT

The authors would like to thank all reviewers who provided constructive feedback on this paper.

REFERENCES

[1] E.PadmaLahari, D.V.N.Siva Kumar,S. Shiva Prasad, "Automatic Text Summarization with Statistical and Linguistic Features using Successive Thresholds", IEEE International Conference on Advanced Communication Control and Computing Technologies, ISBN No. 978-1-4799-3914-5/14 **©2014**.

[2] Alok Ranjan Pal, Diganta Saha, "An Approach to Automatic Text Summarization using WordNet", IEEE International Conference on Advanced Communication Control and Computing Technologies, 978-1-4799-2572-8/14©**2014**.

[3] Ms.Pallavi D.Patil, Prof.N.J.Kulkarni, "Text Summarization Using Fuzzy Logic", International Journal of Innovative Research in Advanced Engineering (IJIRAE) Volume **1** Issue **3**, May **2014**.

[4] D.Y. Sakhare, Dr. Raj Kumar, "Syntactic and Sentence Feature Based Hybrid Approach for Text Summarization", I.J. Information Technology and Computer Science, **03**,Page no- **38-46** , **2014**

[5] A.R.Kulkarni, S.S.Apte, "An Automatic Text Summarization using lexical cohesion and correlation of sentences", International Journal of Research in Engineering and Technology, Volume: **03** Issue: **06** , Jun-**2014**.

[6] Atif Khan, Naomie Salim, "A Review On Abstractive Summarization Methods", Journal of Theoretical and Applied Information Technology. Vol. **59**, No.**1,**10th January **2014**

[7] Dipti Y. Sakhare, Dr.Rajkumar, "Neural Network Based Approach To Study The Effect Of Feature Selection On Document Summarization", ISSN: 0975-4024 Vol. **5,** No. **3**, Jun-Jul **2013**.

[8] Anjali R. Deshpande, Lobo L. M. R. J., "Text Summarization using Clustering Technique", International Journal of Engineering Trends and Technology (IJETT) - Vol.**4** ,Issue**8**- August **2013**.

[9] Mohsen Pourvali and Mohammad Saniee Abadeh, "Automated Text Summarization Base on Lexical Chain and graph Using of WordNet and Wikipedia Knowledge Base", IJCSI International Journal of Computer Science Issues, Vol. **9**, Issue **1**, No **3**, January **2012**.

[10] Manisha Prabhakar, Nidhi Chandra, " Automatic Text Summarization Based On Pragmatic Analysis", International Journal of Scientific and Research Publications, Volume **2**, Issue **5**, May **2012**.

[11] Ms.Meghana.N.Ingole, Mrs.M.S.Bewoor, Mr.S.H.Patil, "Text Summarization using Expectation Maximization Clustering Algorithm", International Journal of Engineering Research and Applications (IJERA) Vol. **2**, Issue **4**, July-August **2012**, pp.**168-171**.

[12] G.PadmaPriya, K.Duraiswamy, "An Approach for Concept-based Automatic Multi-Document Summarization using Machine Learning", International Journal of Applied Information Systems (IJAIS) – ISSN : 2249-0868 ,Volume**3**, No**3**., July **2012**.