Volume 6, Issue 4, April 2016 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

A Survey on Extractive Text Summarization

Richa Sharma, Prachi Sharma
M.Tech Scholar, Department of Computer Science, Banasthali University,
Rajasthan, India

Abstract— In the present scenario of fast growing usage of the Internet, the task of searching for the user becomes very tedious and time consuming task as there are large numbers of results shown about one topic. Among these results which are the most relevant results for the user is one more difficult task. It is also a very difficult task for a user to manually summarize the large text available on the Internet. Text Summarization came into being in order to reduce these difficulties for the user. Text- summarization is the process of reducing the huge amount of data which is available on Internet or on any other source of information like newspapers, books etc., into summarized documents, in order to create the summary of the document that retains the most important points and the overall meaning of the document. The process of text summarization is further divided into extractive text summarization and abstractive text summarization. In Extractive summarization, the summary is generated by using most important sentences from the document exactly as they appear in it and then concatenating them in shorter form. In abstractive text summarization the concept of the original text is understood and then the summary is written in own words without changing the meaning of the original text. This survey paper presents the various techniques used in extractive summarization method.

Keywords— Text Summarization, Extractive Summarization, Natural language processing, Summary Evaluation, Abstractive Summarization

I. INTRODUCTION

Text summarization [1] [2] is a process to reduce the large text document without changing the original meaning of the document. That means, the process of summarization creates a summary [3] of text file or text document, the summary generated not only helps the user to find the relevant information of his/her query but it also gives an overview of written text. With the help of text summarization process less effort and less time of the human being is used. Due to fast growing of Internet, there is high availability of the information, but it is very difficult for a human to find the relevant information from the huge pile of both relevant and irrelevant information [4]. Therefore, there is a need of text summarization. In order to save the time and the human efforts the system is built which automatically retrieves the document from the huge pile of information, categorize it according to the information need of the user and summarize the document as per user's requirements.

Text summarization is of two types which are- Manual summarization and Automatic summarization. Manual summarization [6] is the process through which the summary of the large document is created manually i.e. with the help of human experts, which is a very difficult task for an human expert to summarize the large document manually as deep learning is required for it. Manual summarization have mainly two disadvantages, firstly, the relevant document has to be searched from the Internet where large number of documents on the same topic is available and secondly this whole procedure of manual summarization is time consuming and may be stressful for the human expert. Due to these disadvantages of Manual Text Summarization, Automatic Text Summarization [5] came into being and in the present scenario only automatic text summarization is used. Automatic summarization is the process of creating a summary of large document available on the Internet automatically with the help of a computer program without affecting the overall meaning of the original document.

Text Summarization can be done by two techniques [6] which are Extractive Technique and Abstractive Technique. In Extractive summarization [6] [7], the summary is generated by using most important sentences from the document exactly as they appear in it and then integrated in shorter form. The summary generated by the extractive technique contains those sentences which have highest scores among all the sentences of the original document. The importance of sentences is decided on the basis of statistical and linguistic features of sentences [6].Extractive summarization can be done [8] [9] by following steps:-1) Pre-processing step and 2) Processing step.

In Pre-Processing step [6] [9], the original text is represented in structured form. It includes three sub-processes which are: a) Sentences Segmentation [9] - In this process, sentences in the original text is segmented with the help of punctuation marks like '.', '?'. '!' etc. b) Stop-Word Removal [9]-Common words with no meaning like 'is', 'am', 'a', 'the', 'in' etc. are removed in this step as these words do not play an important role in the summary generation. c) Stemming [9] [10] [11]-In this, words in the text are stemmed into their root word respectively.

In processing step, the various features are decided and calculated to score the sentence. After that using weight learning method [6] [9], the weights are assigned to each feature. Now final score of each sentence is calculated using feature weight equation and those sentences are chosen for the final summary whose scores are highest among all the sentences in the original text.

In Abstractive Technique [6] [10], the main concept of the document is understood then those concepts are expressed in precise form using simple natural language. In this technique the generated summary doesn't include same sentences from input text as compared to the extractive technique.

Summary evaluation [19] [20] is very important aspect for text summarization. Generally, the summaries can be evaluated by extrinsic method [20] which attempts to evaluate the summary on the basis of information retrieval oriented task. For evaluation of summary precision and recall is used. Intrinsic method [20] attempts to evaluate summary with the help of human expert, which is also known as gold summary.

II. LITERATURE SURVEY

Vishal Gupta and Gurpreet Singh Lehal, "A Survey of Text Summarization Extractive techniques". In this paper author describes the extractive summarization methods which comprises of two parts Pre Processing and Processing. In this paper, pre-processing step is further divide into other sub processes which are sentence segmentation, stop word removal and stemming. In processing step, the weights are given to the features used for extraction of summary from the large document respectively [6].

Saranyamol C S and Sindhu L, "A Survey on Automatic Text Summarization." In this paper the author describes about the various techniques used in automatic text summarization which are extractive text summarization and abstractive text summarization respectively [7].

Rafael Ferreira, Luciano de Souza Cabrala, Rafael Dueire Lins , Gabriel Pereira Silva , Fred Freitas , George D.C. Cavalcanti , Rinaldo Lima a, Steven J. Simske , Luciano Favaro, "Assessing sentence scoring techniques for extractive text summarization." This paper, gives the brief description of various features used to perform extractive summarization and it also describes the methods for summary evaluation [1].

K. Vimal Kumar, Divakar Yadav "An Improvised Extractive Approach for Hindi Text Summarization." This paper mainly laid emphasis most importantly on the Hindi text summarization. It also describes various features used for the Hindi summarization using extractive approach of text summarization. The author had proposed a system which can generate the summary with 85 % accuracy [14].

Vishal Gupta, "Hybrid Algorithm for Multilingual Summarization of Hindi and Punjabi Documents." The author of this paper has proposed a hybrid algorithm for Hindi and Punjabi text summarization. The algorithm proposed by the author is the first algorithm which can summarize both Hindi as well as Punjabi text. It also suggests some new methods for Hindi and Punjabi text [15].

Ani Nenkova "Summarization Evaluation for Text and Speech: Issues and Approaches." This paper suggests the methods for summary evaluation after the process of text summarization. Also, it describes some human models for summary evaluation [16]**.**

Inderjeet Mani "Summarization Evaluation: An Overview." The author describes various methods for evaluating summary. It also describes the various advantages of summary evaluation respectively [17].

III. VARIOUS FEATURES TO SCORE A SENTENCE

There are several features which are used to score the sentences, some of them are as follows:

A. Cue-Phrase Feature

Phrases such as "In summary", "It concludes", "Finally", and "In conclusion" etc. [15] are defined as cue phrases. If the sentence starts with any of the cue phrase then that sentence is considered as important sentence than others because these phrases indicates the resulting information about the text.

B. Title Word Similarity Feature

Title of document defines the overall theme of document that's why sentences containing words which are similar or ought to be similar to the title of the document are most likely to be included in the summary [6].

C. Sentence Length Feature

Lengthy sentences are preferred to be included in summary as compared with short sentences because usually short sentences are less informative [15].

D. Proper Noun Feature

The name of the person (Kamal, Neeraj, Sudha, and Narendra Modi etc.), place (New Delhi, Mumbai, Shyam Nagar etc.), thing (Cello pen, Apple computer etc.) are considered as the proper nouns. So, the sentences having large number of proper nouns are considered as important for summary [6] [18].

E. Font Feature for Sentences

Sentences which are having words in bold, italics or underlined or having font size greater than rest of text are considered as important and should be included in summary [15].

F. Upper Case Feature

The sentences of the input text containing words starting from the capital letters are considered as important for the summary. This feature is not applicable for Hindi text summarization [6].

G. Sentence Position Feature

The sentence position is referred to as the position of the sentences in the document. Usually first sentence of first paragraph and last sentence of last paragraph are considered as important because sentence in beginning shows the theme of document and last sentence conclude the document. So sentences at these position are likely to be included in summary [18].

H. Numeric Data Feature

The sentences containing numeric data like 1, 2, 3 and i, ii, iii….etc. or the sentences containing information about the money, profit, loss, damage, etc. are considered as important and are included in the summary [15] [18].

I. Average IF-ISF (Term Frequency Inverse Sentence Frequency) Feature

TF-ISF is used to calculate the importance of a term in whole document. The importance increases as number of times a term appears in a sentence (TF) but is balanced by the frequency of the term in the document (ISF) [18].

J. Sentence Centrality Feature

In this for each sentence s, similarity is calculated with respect of other sentences s'. At last the similarities are add up in order to find the value of this feature for sentence s [6].

K. Pronoun Feature

Sentences comprising pronoun such as 'he', 'she', 'it', 'they', 'these', 'those' and 'their' etc. should not be included in summary without expansion to their corresponding nouns [6]. Because sentences in summary containing only pronoun with absence of their corresponding noun wouldn't be self-explanatory.

L. Non-Essential Information Feature

Sentences containing words in beginning such as 'because', 'furthermore', 'typically', 'additionally' etc. should not be included in summary because those sentences usually indicates non-essential information [6] [15].

IV. TECHNIQUES USED FOR EXTRACTIVE TEXT SUMMARIZATION

There are several techniques used to generate extractive summary.

A. Term Frequency-Inverse Document Frequency (TF-IDF) Method

This method uses the TF/IDF score of sentence in order to generate the final summary. In this, the sentence frequency i.e. the number of sentences in the document that contain a particular term is calculated. Now similarity is calculated between query and these sentence vectors and highest scoring sentences are included in summary. This summarization is query specific, to produce generic summary first remove stop-words then calculate frequency of each remaining term i.e. TF. Now high frequency words in the document may be taken as query words [6] [7].

B. Cluster Based Method

Normally, if documents are written for different topics, they are divided into sections either implicitly or explicitly to generate a significant summary. This aspect is known as clustering. The overall score of a sentence is calculated as weighted sum of three factors such as: similarity of sentence to the theme of a particular cluster, location of sentence in the document and similarity of the sentence to the first sentence in the document to which it belongs [7].

C. Graph Theoretic Approach

This method is basically used to determine the theme of the passage. After the pre-processing steps, sentences of the original document are represented as nodes in an undirected graph. If any of the sentences share common words, then the nodes in the graph are connected with common edge. The edges with the high cardinality are important sentences are likely to be included in the final summary of the document [6].

D. Machine Learning Method

The machine learning method uses the set of training document and their extractive summaries in order to generate the summary of the input text. This technique shapes the summarization process as a classification problem. In this, the sentences of the input text are categorized in two classes as summary and non-summary sentences. The classification of sentences are learnt from training data with the help of features used to score the sentences and Bayes' rule [24].

E. LSA Method

LSA stands for Latent Semantic Analysis. In this method, Singular value decomposition applies to document-word matrices, group of documents that are associated to each other conceptually even if they don't have a common word. Advantage of LSA vectors over word vectors is that semantic relations as represented in human brain are captured automatically by LSA vectors while word vectors require explicit methods to originate those semantic relations [25].

F. Text Summarization with Neural Networks

This technique accomplishes its work in 2 phases. First phase includes training of neural network to learn whether the sentence from test paragraph should be included in summary or not. This process is done by human reader [22]. During

this training neural network learns about features that are occurred in summary sentences. After getting the feature information, there is need to identify the trends and relationships between the features that are inherent in the majority of the sentences. This is done by second phase, called feature fusion phase [23]. This phase completes in 2 steps: 1) eliminating uncommon features and 2) collapsing the effects of common features.

G. Text Summarization Based on Fuzzy Logic

In this method all features such as sentence length, similarity to title, similarity to key word etc. are taken as input to fuzzy system [5] [21]. This system also consists a knowledge base of all IF-THEN rules required for summarization. According to the features and rules applicable for a sentence, a value ranging from 0 to 1 is determined for each sentence. Now, the obtained value determines the rank of sentence for final summary.

H. Query Based Extractive Text Summarization

In this method, the sentences are scored according to the word frequency. The query is given as the input to the system and those sentences which are containing query phrases, given higher scores as compared to the sentences containing single query words. At last the sentences of the input text having highest scores among all the sentences are incorporated in the final summary. Sentences can also be mined from different sections and subsections. In this method number of summary sentences with their extended context are depends upon size of output screen that can be seen without scrolling [6] [7].

V. CONCLUSION

This survey paper gives the details on extractive text summarization features and its methods. The extractive text summarization is a process of selecting important sentences from the document and including those sentences as it is in the final summary of the document and the selection procedure of sentences is done on the basis of statistical and linguistic features of the sentences. Without use of NLP extractive summary may suffer from lack of cohesion and semantic.

The text summarization software should generate summary in less time and with least redundancy. There are two approaches for evaluating summary- intrinsic methods or extrinsic methods.

REFERENCES

[1] Rafael Ferreira et al. "Assessing Sentence Scoring Techniques for Extractive Text Summarization", Elsevier Ltd., Expert Systems with Applications 40 (2013) 5755-5764.

[2] D. Das and A. F. Martins, "A Survey on Automatic Text Summarization", Literature Survey for the Language and Statistics II course at CMU, Vol. 4, pp. 192-195, 2007.

[3] M. Haque *et al.* "Literature Review of Automatic Multiple Documents Text Summarization", International Journal of Innovation and Applied Studies, Vol. 3, pp. 121-129, 2013.

[4] H. P. Luhn, "The automatic creation of literature abstracts", IBM Journal of Research and Development, vol. 2, pp. 159-165, 1958.

[5] Kyoomarsi F., Khosravi H., Eslami E., Dehkordy P.K., "Optimizing Text Summarization Based on Fuzzy Logic." In proceedings of Seventh IEEE International Conference on Computer and Information Science, IEEE, University of Shahid Bahonar Kerman, UK, 347-352, 2008.

[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.

[7] Saranyamol C S and Sindhu L, "A Survey on Automatic Text Summarization", International Journal of Computer Science and Information Technologies, Vol. 5(6), pp. 7889-7893, 2014.

[8] Vishal Gupta and G.S Lehal, "A Survey of Text Mining Techniques and Applications", Journal of Emerging Technologies in Web Intelligence, Vol. 1, No. 1, pp. 60-76, August 2009.

[9] Neelima Bhatia and Arunima Jaiswal, "Trends in Extractive and Abstractive Techniques in Text Summarization", International Journal of Computer Application (0975-8887), Vol. 117- No. 6, May 2015.

[10] V. Gupta and G.S. Lehal, "A Survey of Common Stemming Techniques and Existing Stemmers for Indian Languages," Journal of Emerging Technologies in Web Intelligence, Vol. 5, pp. 157-161, 2013.

[11] Vishal Gupta, "Hindi Rule Based Stemmer for Nouns", International Journal of Advanced Research in Computer Science and Software Engineering, Vol. 4, Issue 1, January 2014.

[12] N. R. Kasture, Neha Yargal, Nityanand Singh, Neha Kulkarni, Vijay Mathur, "A Survey Methods of Abstractive Text Summarization", International Journal for Research in Emerging Science and Technology, Vol. 1, Issue 6, November 2014.

[13] Atif Khan and Naomie Salim, "A Review on Abstractive Summarization Methods", Journal of Theoretical and Applied Information Technology, Vol. 59, No. 1, January 2014.

[14] K. Vimal Kumar, Divakar Yadav "An Improvised Extractive Approach for Hindi Text Summarization" Springer India 2015, J.K. Mandal et al. (eds.), Information Systems Design and Intelligent Applications, Advances in Intelligent System and Computing 339, DOI 10.1007/978-81-322-2250-7_28.

[15] Vishal Gupta*," Hybrid Algorithm for Multilingual Summarization of Hindi and Punjabi Documents"* 2013 in Springer International publishing Switzerland 2013.

[16] Ani Nenkova, "Summarization Evaluation for Text and Speech: Issues and Approaches", Stanford University.

[17] Inderjeet Mani, "Summarization Evaluation: An Overview", USA.
[18] Deepali P Kadam, et al., "A Comparative Study of Hindi Text Summarization Techniques: Genetic Algorithm and Neural Networks." International Journal of Innovation and Advancement in Computer Science, vol. 4, ISSN 2347-8616, March 2015.
[19] Chin-yew Lin, "A package for automatic evaluation of summaries", in Proc. ACL Workshop on text summarization branches out, 2004.
[20] Ani Nenkovaand, Rebecca Passonneau, "Evaluating content selection in summarization: The Pyramid method", in HLT-NAACL, pp. 145-152, 2004.
[21] Ladda Suanamali, Naomic Salim, Mohammed Salem and Binwahlan, "Sentence Features Fusion for Text Summarization using Fuzzy Logic", IEEE, 142-145, 2009.
[22] Khosrow Kaikhah, "Automatic Text Summarization with Neural Networks", In Proceedings of Second International Conference on Intelligent Systems, IEEE, 40-44, Texas, USA, June 2004.
[23] Khosrow Kaikhah, "Text Summarization using Neural Networks", Department of faculty publications-Computer Science, Texas State University, eCommons, 2004.
[24] Joel larocca Neto et al., "Automatic Text Summarization using Machine Learning Approach", Book: Advances in Artificial Intelligence: Lecture Notes in Computer Science, Springer Berlin / Heidelberg, Vol. 2507/2002, 205-215, 2002.
[25] Madhavi K. Ganapathiraju, "Overview of Summarization Methods", 11-742, Self-paced lab in Information Retrieval, November 26, 2002.

