A Survey of Automatic Text Summarization Techniques for Indian and Foreign Languages

Prachi Shah
Department of Information Technology
Dharmsinh Desai University
Nadiad, India
prachishah2401@gmail.com

Nikita P. Desai
Department of Information Technology
Dharmsinh Desai University
Nadiad, India
npd_ddit@yahoo.co

Abstract— **Today in the era of Big Data, textual data is rapidly growing and is available in many different languages. In the fast-moving world, it's difficult to read all the text-content. Hence, the need for text summarization is being in the spotlight. Automatic text summarization is a technique which compresses large text to a shorter text which includes the important information. There are two types of summaries: Extractive summaries and Abstractive summaries. Extractive summaries are produced by extracting the whole sentences from the source text. Abstractive summaries are produced by reformulating sentences of the source text. Several text summarization techniques have been proposed in past years for English and various European languages but there are very few techniques that can be found for native languages of India. This paper presents a survey of text summarization techniques for various Indian and foreign languages like English, European, etc. Also, an approach for summarizing Hindi text using machine learning technique has been proposed. We have also described few challenges which are still under research.**

Keywords—text summarization; text mining; extraction; summary genration;

I. Introduction

Automatic text summarization is the technique which compresses a large text to a shorter text which includes the important information. The computer program is given a text and it returns a summary of the original text. This is done by reducing redundancy of the text and by extracting the essence of the text. The technique has been developed through research for more than 50 years [1] and with the extensive use of the Internet and the creation of websites and online textual resources, the need for fast and reliable text summarization is being in the spotlight. However, the ability of a human to understand and organize the large documents is limited. To read long documents consumes valuable time in understanding the essence of the document. Summarization is a well-known technique which solves such problems most accurately.

The output of summary can be of two types: Extractive summaries and abstractive summaries. Extractive summaries are produced by extracting the whole sentences from the source text. The importance of sentences is determined based on statistical and linguistic features of sentences.

Abstractive summaries are produced by reformulating sentences of the source text. An Abstractive summarizers [21][22] understands the main concepts in a document and then convey those concepts in clear natural language. It uses linguistic methods to examine and interpret the text and then to find the new concepts and terms to best describe it by generating new shorter text that conveys the most significant information from the original text document [18].

This paper emphasizes on survey and performance analysis of automatic text summarizers for various Indian and foreign languages.

II. Text Summarizers For Foreign Language

Different text summarization techniques for various foreign languages are discussed below:

A. English

H.p.Luhn [2] is the father of automatic text summarization. He started creating an abstract of English technical literature by automatic means to get quick and accurate identification of technical papers. The main goal of Luhn was to save readers effort and time in finding relevant information for articles and reports. The whole text of a document was prepared in machine readable form and it was scanned by IBM 704 machine and after processing the statistical information is derived. Word frequency and distribution was used for computing relative measures, first for each word and secondly for sentences.

B. Chinese

Po Hu, Tingting He, and Donghong Ji [3] in 2004 proposed a special technique that produces text summary by detecting thematic areas in Chinese documents. In this system, the detection of latent thematic areas is realized by adopting the method of K-medoids clustering as well as a novel method of clustering analysis, which can be used to find out automatically K, the number of clusters. Additionally, a novel parameter, known as representation entropy, has been used for summarization redundancy evaluation. 30 documents of different genres from the Modern Chinese Corpus of State Language Commission were used for testing and its evaluation performance resulted with 68% accuracy for Economic documents.

Yu, Lei, Mengge Liu, Fuji Ren, and Shingo Kuroiwa [4] in 2006 proposed a method for Chinese text, which collects original news text from on-line sources and extracts sentences from them automatically to generate a summary. Based on this, they adopted WML(Wireless Markup

978-1-4673-9939-5/16/$31.00 ©2016 IEEE

Language) to construct the news website for mobile devices which browsing through the news summary. The system is mainly prepared by Automatic News Collection and Auto Text Summarization. The system is based on statistical information and the structural information of the text. The importance of sentences is calculated using noun phrases or noun clauses and importance of words is calculated based on TF-IDF. Also, sentence position is considered as an important feature. 30 Chinese news texts were used to construct the testing corpus. Three students were asked to make summary manually for comparing the system generated summary. For each text, the precision and recall are computed and resulted in the average scores 0.74 and 0.76 at 20% compression ratio.

C. *Arabic*

Sobh Ibrahim, Nevin Darwish, and Magda Fayek [5] in 2006 introduced a trainable Bayesian approach for Arabic extractive text summarization. The features used are TFIDF, sentence length, sentence position, sentence paragraph position, paragraph length. The trainability feature of the system makes it feasible to be customized for a specific domain. System performance has overcome four ad-hoc systems. The performance of system increases when combining sentence weight, sentence length, and sentence position. Adding other features like sentence paragraph position and sentence paragraph order resulted in slight change in performance of the system, because of the fact that most of the paragraphs in the documents have length of only two or three sentences. System is evaluated in terms of precision, recall and F-measure and the system average precision is 68.07%.

El-Shishtawy, Tarek, and Fatma El-Ghannam [6] in 2012 described an efficient generic summarization algorithm for Arabic texts based on extractive summarization approach. Important key phrases of the document for summarization are identified by employing combinations of different linguistic and statistical features. The sentence extraction algorithm uses key phrases as the most important attributes to rank a sentence. The average result of similarity between the extractor and the proposed system is nearly 66% at 25% compression ratio.

D. *Turkish*

Kutlu, Mücahid, Celal Cığır, and Ilyas Cicekli [7] in 2010 proposed the first generic text summarization technique for the Turkish language. In order to extract sentences to form a summary with coverage of the key content of the text and less redundancy, surface level features are used such as term frequency, centrality, title similarity, key phrase, and sentence position. The rank of the sentence is calculated using a score function which uses its feature values and the weights of the features. Feature weights are learned using a machine learning technique with the help of human-generated summaries. Performance evaluation is carried out by comparing output summary with manual summaries of two newly created Turkish data sets.

E. *Swedish*

Gustavsson and Arne J¨onsson [8] have presented results from evaluations of an automatic text summarization technique that uses a combination of Random Indexing and PageRank. In experiments, they have used two types of texts: newspaper texts and government texts. The result shows that text types, as well as other aspects of texts of the same type influence the performance. Combining Random Indexing and PageRank provides the best results on government texts.

III. Text Summarizers For Indian languages

Different text summarization techniques for various Indian languages are discussed below:

A. *Tamil*

Sankar K, Vijay Sundar Ram R and Sobha Lalitha Devi [9] in 2011 proposed summarization system for Tamil language based on scoring of sentences, in summary, using graph theoretic scoring technique. The system uses statistics of frequency of words and a position of term and calculation of sentence weight by using string pattern for ranking sentences. This text ranking algorithm is not a domain specific and also do not need any annotated corpora. They have used ROUGE evaluation toolkit to evaluate the proposed algorithm and the average Rouge score is 0.4723.

Banu, Karthika, Sudarmani and Geetha [10] in 2007 proposed summarizer for Tamil documents which extracts sentences from a single document using sub graph and generates a generic document summary. In this system, Language-Neutral Syntax (LNS) has been used for considering the semantics of the document. It also uses syntactic analysis of the original text which analyzes logical form used in each and every sentence. Triples of Subject Object Predicate are chosen from each sentence to generate a semantic graph of an original document and its corresponding summary extracted by human experts. The Support Vector Machine (SVM) classifier has been used for training, to identify SOP triples from the document semantic graph which belongs to the summary. Using this classifier, it extracts automatic summaries from the test documents.

B. *Kannada*

Jayashree.R, Srikanta Murthy and Sunny.K [11] in 2011 presented a method to generate extractive summaries of documents in the Kannada language. This proposed algorithm extracts keywords from pre-categorized Kannada documents collected from online resources. The combination of GSS (Galavotti, Sebastiani, Simi) coefficients and IDF (Inverse Document Frequency) methods all along with TF (Term Frequency) has been used for extracting keywords and later used these for summarization. A document from a given category was selected from the database and depending on the number of sentences specified by the user, a summary was generated. The result of a manual evaluation of the summarizer with three different human summaries among various categories sports, Entertainment, literature gives average recall value– 0.76,– 0.8– 0.7.

J. S. Kallimani, K.G. Srinivasa and B. R. Eswara [12] in 2010 proposed a text summarizer for Kannada language i.e.

"AutoSum", which is a named IR system using Text Summarization of some regional Languages in India. This system processes the input text and then decides which sentences are significant and which sentences are not significant. The output summary of this system can be created either in plain text or in HTML. If HTML is used in output then significant sentences are highlighted. This summarizer follows up 3 main steps:

1) User gives command on the terminal

2) Next, the input processes through the system and summary is generated

3) The resulting text is sent to the terminal after summary is generated or the results of the summary are highlighted in the web browser.

This system makes use of nouns, adjectives and adverbs as key terms. The value of the score of each feature of each sentence is determined and summed up to the score of that sentence. Every sentence is assigned a score based on the key features in it [19].

C. Bengali

Kamal Sarkar [13] proposed a technique for Bengali text summarization using sentence extraction technique. The approach proposed here has three major steps: (1) preprocessing (2) sentence ranking (3) summary generation. The preprocessing step includes removal of stop-word, stemming and segmenting the input document into a collection of sentences and then sentences are given rank using a few features such as thematic term, position value. The thematic terms are the terms which relate to the main theme of a document. The score of a sentence position is calculated in which the first sentence of a document gets the highest score and the last sentence gets the lowest score. Long sentences are given more preference. A summary is generated after giving rank to the sentences based on their scores and selecting topmost ranked sentences. To test summarization system, 38 Bengali documents from the Bengali daily newspaper were used. The result of evaluation of summary with single reference summary gives average unigram based recall score - 0.4122.

A. Das and S. Bandyopadhyay [14] in 2010 developed opinion text summarizer for Bengali language based on the given topic which can find out the information on sentiments in the source text. A model is applied on topic-sentiment for determination and aggregation of sentiments. It is implemented for theme determination. Furthermore, aggregation is performed by clustering theme using k-means approach and also by applying theme graph representation, which is at last applied for selecting relevant sentences in summary by using standard page rank approach. This summarization system evaluated result with Precision of 72.15%, Recall of 67.32% and F-measure of 69.65%.

D. Bangala

Mohammad Ibrahim, Humayun Kayesh [15] in 2013 developed an extraction based summarization technique which works on the text documents of Bangala language. The system summarizes a single document at a time. In the summarization process, the countable features like word frequency, sentence positional value, and cue words were used. They have used 45 Bangla news articles to test their system. The proposed technique has been compared with the summary of documents which is generated by human experts. The evaluation of the system shows that 83.57% of summary sentences selected by the system agreed with those made by the human.

E. Punjabi

Visual Gupta and Gurpreet Singh Lehal [16] in 2013 proposed an extractive automatic text summarization system for the Punjabi language. They have described the system which consists of two stages 1) Pre-Processing and 2) Processing, where pre-processing is defined as the stage which identifies the sentence boundary, word frequency, eliminates Punjabi stop words etc. and in the processing stage sentence features are calculated and a weight is assigned to each sentence. The author tested the proposed system over 50 Punjabi news documents and 50 Punjabi stories. The accuracy of the system varies from 81% to 92 %.

F. Hindi

K. Vimal Kumar and Divakar Yadav [17] in 2015 proposed an improvised extractive approach to Hindi text summarization. The system was based on an algorithm for scoring the sentences based on co-occurrence of the radix of thematic words. The average accuracy of the system with the expert's manual summary found to be 85 %.

IV. Proposed Methodology

Numerous research projects are investigating and exploring various techniques of automatic text summarization system for the English and European languages and also for Asian languages such as Chinese, Japanese, etc. However, we have analyzed that less research work has been done for Indian languages. Though Hindi is the top-most language used in India and also in a few neighboring countries there is a lack of proper summarization system for Hindi text.

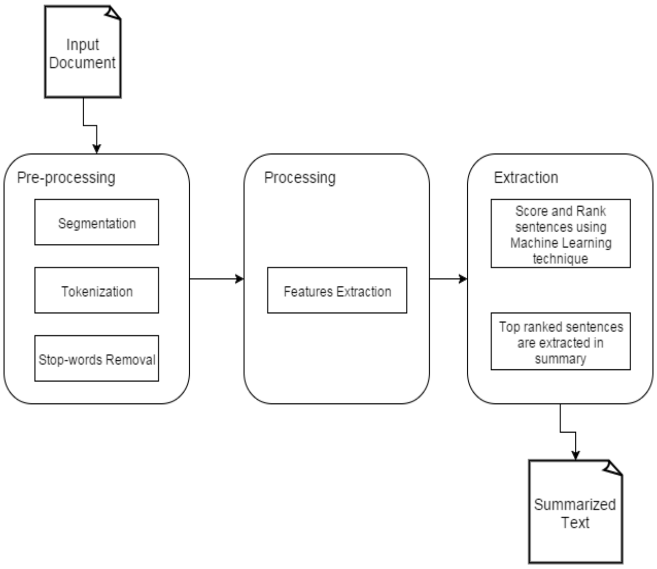

Fig. 1. Workflow of proposed methodology

Based on the analysis it is found that it has become necessary to come up with ways in which the different indicators can be combined. [23] Machine learning is a technique in which we can combine various indicators and can decide what features should be used and how they should be weighted relative to each other.

Here, an approach for automatic Hindi text summarization using machine learning technique has been proposed. The workflow of this system is as shown in fig.1. The system consists of 3 major blocks: pre-processing, processing and extraction. The pre-processing phase includes segmentation, tokenization, and stopwords removal. The processing is one of the most important phases of text summarization. It includes feature extraction and for this system, features to be used are sentence position, sentence length, numerical data, presence of inverted comma and keywords in the sentence. In the extraction phase, a machine learning technique is applied to identify whether the sentence should be included in summary or not based on the training set. Then the sentence is given rank based on sentence score. Top-ranked sentences are considered in the summary.

Based on this proposed method, we have done some experiments with approximately 30 Hindi texts collected from online sources. For initial testing, we have manually calculated the score for all the five features for every sentence and stored in an excel file. We have used Lib-SVM a machine learning tool to tag a sentence whether it should be in summary (1) or not in summary (0). About 100 sentences were taken as a training set to train a model. Other remaining sentences were tested using that trained model. Initially, we have achieved 75% accurate result. 18 sentences were correctly classified from 24 sentences.

V. Conclusion

In this paper, a brief summary of automatic text summarization techniques for various Indian and Foreign languages has been described. We can notice that good work has been done for various foreign languages like English, Turkish, Arabic, etc. But automatic summarization system for Indian languages is still lacking. We can also conclude that different combination of features works differently for different types of content. Hence, it is challenging to create a single summarizer for different types of content.

In future, we are aiming to use more features for extracting Hindi sentences. Also, we will try different machine learning techniques for comparison and try to achieve better accurate results. Also, try to test our technique rigorously on large dataset of various domains like news, autobiography, etc.

References

[1] Lloret E., “Text summarization: an overview” Paper supported by the Spanish Government under the project TEXT-MESS (TIN2006-15265-C06-01). 2008.

[2] Hans Peter Luhn. “The automatic creation of literature abstracts,” *IBM Journal of research and development*, 2(2):159–165, 1958.

[3] Po Hu, Tingting He, and Donghong Ji. "Chinese text summarization based on thematic area detection." In ACL-04 Workshop: Text Summarization Branches Out, Barcelona, Spain, Association for Computational Linguistics, pp. 112-119. 2004.

[4] Yu Lei, Mengge Liu, Fuji Ren, and Shingo Kuroiwa, "A Chinese Automatic Text Summarization system for mobile devices." In The Pacific Asia Conference on Language, Information and Computation (PACLIC-2006), pp. 426-429. 2006.

[5] Sobh Ibrahim, Nevin Darwish, and Magda Fayek, "A trainable Arabic Bayesian extractive generic text summarizer." In Proceedings of the Sixth Conference on Language Engineering ESLEC, pp. 49-154. 2006.

[6] El-Shishtawy, Tarek, and Fatma El-Ghannam, "Keyphrase based Arabic summarizer (KPAS)." In 8th International Conference on Informatics and Systems (INFOS), pp. NLP-7. IEEE, 2012.

[7] M. Kutlu, C. Cigir, and I. Cicekli, “Generic text summarization for Turkish.” The Computer Journal, vol.53, no.8, pp.1315-1323,2010.

[8] Gustavsson, Pär, and Arne Jönsson. "Text summarization using random indexing and pagerank." Proceedings of the third Swedish Language Technology Conference (SLTC-2010), Linköping, Sweden. 2010.

[9] S. Kumar, V. S. Ram and S. L. Devi, “Text Extraction for an Agglutinative Language,” Proceedings of Journal: Language in India, pp. 56-59, 2011.

[10] M. Banu, C. Karthika, P Sudarmani and T.V. Geetha,“Tamil Document Summarization Using Semantic Graph Method”, Proceedings of International Conference on Computational Intelligence and Multimedia Applications, pp. 128-134, 2007.

[11] R. Jayashree, K. M. Srikanta and K. Sunny, “Document Summarization in Kannada using Keyword Extraction,” Proceedings of AIAA 2011,CS & IT 03, pp. 121–127 , 2011.

[12] J.S. Kallimani, K.G. Srinivasa and B. R. Eswara, “Information Retrieval by Text Summarization for an Indian Regional Language,” In Proceedings of International Conference on Natural Language Processing and Knowledge Engineering, pp. 1-4, 2010.

[13] K. Sarkar, “Bengali text summarization by sentence extraction,” In Proceedings of International Conference on Business and Information Management (ICBIM-2012), NIT Durgapur, pp. 233-245, 2012.

[14] A. Das and S. Bandyopadhyay, “Topic-Based Bengali Opinion Summarization”, International Conference COILING ’10, Beijing, pp. 232–240, 2010.

[15] Efat, Md Iftekharul Alam, Mohammad Ibrahim, and Humayun Kayesh. "Automated Bangla text summarization by sentence scoring and ranking." In International Conference on Informatics, Electronics & Vision (ICIEV), , pp. 1-5, IEEE, 2013.

[16] Gupta, Vishal, and Gurpreet Singh Lehal. "Automatic Text Summarization System for Punjabi Language." Journal of Emerging Technologies in Web Intelligence, pp. 257-271, 2013.

[17] Kumar, K. Vimal, and Divakar Yadav. “An Improvised Extractive Approach to Hindi Text Summarization.” In Information Systems Design and Intelligent Applications, pp. 291-300, Springer India, 2015.

[18] Gupta Vishal and Gurpreet Singh Lehal. “A survey of text summarization extractive techniques.” Journal of Emerging Technologies in Web Intelligence, vol. 3, pp.258-268,2010.

[19] Gupta Vishal. "A Survey of Text Summarizers for Indian Languages and Comparison of their Performance." Journal of Emerging Technologies in Web Intelligence vol.5,no. 4,pp.361-366,2013.

[20] Dhanya, P. M., and M. Jathavedan, "Comparative Study of Text Summarization in Indian Languages." International Journal of Computer Applications vol.75, no.6,2013.

[21] G Erkan and Dragomir R. Radev, “LexRank: Graph-based Centrality as Salience in Text Summarization”, Journal of Artificial Intelligence Research, Re-search, Vol. 22, pp.457-479,2004.

[22] Udo Hahn and Martin Romacker, “The SYNDIKATE text Knowledge base generator”, Proceedings of the first International conference on Human language technology research, Association for Computational Linguistics, ACM, Morristown, NJ, USA, 2001.

[23] Das, Dipanjan and André FT Martins. "A survey on automatic text summarization." Literature Survey for the Language and Statistics II course at CMU 4, pp. 192-195, 2007.