ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 8, August 2016

A Survey of Text Summarization Techniques for Different Indian Regional Languages

Nilofar Mulla, Shital K. Dhamal

M.E Student, Dept. of Computer Engineering, Mumbai University, Maharashtra, India

Dept. of Computer Engineering, Mumbai University, Maharashtra, India

ABSTRACT: Text summarization is the process of extracting needed information from the source text and to present that information to the user in the form of summary. It is very difficult for human beings to summarize large documents of text manually. Automatic summarization provides the required solution as well as challenging task because it requires deep analysis of text. There are two types of summarization: Extractive summarization and Abstractive summarization. The Extractive summaries are produced by extracting the whole sentences from the source text. Abstractive summaries are produced by reformulating sentences of the source text. This paper is about a survey of text summarization techniques for various Indian regional languages like Hindi, Punjabi, Tamil Kannada and Bengali. The proposed system is based on English language text summarization in which Naming entity reorganization and Part Of speech is used for feature extraction and graph is generated for text summarization

KEYWORDS: text summarization; extraction; abstraction; summary generation;

I. Introduction

Text summarization is a technique in which the large data is compressed in such a way that the actual meaning of data doesn’t change. The data arrange in such a way that reader get full knowledge of the large text. The use of electronic information increases day by day, people use internet to search required information. Large amount of data is stored at internet. It is very difficult and time consuming to handle such a large data. It is not possible for user to read whole data, hence text summarization is used to summarize the data, and that summarized data is displayed to the user, so that user easily understands the data. Text summarization is divided into two categories: i) Extraction ii) Abstraction Extractive summaries are generated by reusing portions of text like words, sentences, etc. verbatim. For example, search engines basically generate extractive summaries of WebPages. Most of the time the summarization research today is on extractive summarization. In abstractive summarization, information from the source text is re-generating. Human beings generally write abstractive summaries Abstractive summarization has not properly because allied problems such as semantic representation, inference and natural language generation are relatively harder. The abstractive technique is substitute to original documents rather than part. Abstractive summaries are generated by interpreting main concept of document and then stating that content in natural language. The extractive summaries are generating via removing redundant sentences or words from original text the sentences then concatenate into shorter text to generate meaningful summary.

II. Related work

Hindi:

Chetan Thaokar And Latesh Malik [1] in 2013 proposed an extractive approach to Hindi text summarization, here the linguistic approach is used to find most relevant sentence and perform the steps like pre-processing feature extraction and genetic algorithms for ranking the sentence

K Vimal Kumar And Divakar Yadav [2] in 2015 proposed an enhanced extractive approach to summaries the Hindi text here the score of the sentence is calculated based on occurrence of word the accuracy of the system is 85%.

Kannada:

Varsha R Embar , Surbhi despande (3) in 2013 proposed an abstractive approach to summaries the kannada language text Here the sAramsha system is involve which analyse the document perform the pos (part of speech) in

DOI: 10.15680/IJIRCCE.2016. 0408076

pos it is the words from text is categorized in part in such a way like noun, verb adjective after pos the streaming operation is performed in which stemmers are removed from word then the NER tool which label sequence of words in text which are name of things such as person and company name, organization locations etc IE rules are used to extract the sentences a and then generate the summary.

JayaShree R , Srikanta Murthy and sunny k [4] in 2011 proposed an extract I've approach to summaries the text in Kannada language in this system the keywords are extracted from the Kannada documents here the documents are collected from online resources here the inverse document frequency (IDF) and term frequency term are used for extracting the keywords from text after selecting the keywords the documents from the data base is selected according to selected keywords and summary was generated.

Panjabi :

Vishal Gupta and Gurpreet Singh lehal [5] in 2013 proposed an extractive approach to summaries the Panjabi document with multiple news by rating the sentence based on linguistics features it has two phases
1) Pre-processing phase which Panjabi stop words stemmer for Punjabi nouns and pronoun is removed
2) Processing phase in this case the different features are decided and according to that weight of the sentence is calculate and the sentence for the having large weight is selected for summarization the author test for the earthen tested for 50 Punjabi news documents and for 60 stories in Punjabi to the accuracy of system is in between 81% to 92%.

Tamil:

Sankar K, Vijay Sundar Ram R and Sobha Lalitha Devi [5] in 2011 proposed a system for summarization of Tamil language documents. This system is based on scoring of sentences here, the graph theoretical score technique is used to scoring the sentences and is done by using string patterns. The ranking algorithm is not domain specific and they used rough evaluation tool kit.

Bengali:

Md. Iftekharul Alam Efat Mohammad Ibrahim Humayun Kayesh [7] in 2013 proposed a system for summarization of Bengali text documents summarization here the extractive approach used. Here, first the pre-processing is done means tokenization stop word and stemmers are removed. The second step is sentence ranking and summarization is done .the sentences are ranked based on frequency, position value, cue words and skeleton of document (title and headers) then the data is summarized by using ranking value. The average accuracy of system is 83.57%.

III. Proposed Algorithm

In proposed system text summarization is done on English language documents. In this first pre-processing phase is done in which tokenization, stop word removal and stemmer removal is done. Here improved Stemmer for removed the limitations like better word. We Use NER tool for better Stability to summarization Module. The extractive approach is used. We first finding important word then the sentence based on the words. We represent the summarized data in graph format as well as text format for better representation.

IV. Conclusion and Future Work

In this paper, a brief summary of automatic text summarization techniques for various Indian regional languages has been described. We can notice that good work has been done for various languages like Hindi, Tamil, and Panjabi etc. But automatic summarization system for English languages still needs some enhancement. 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 English sentences. Also, we will try different machine learning techniques for comparison and try to achieve better accurate results. Also, try to use graph based approach for display the end result.

REFERENCES

1. Chetan Thaokar And Latesh Malik " Test Model for Summarizing Hindi Text using Extraction Method"In Proceedings of 2013 IEEE Conference on Information and Communication Technologies (ICT 2013
2. 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
3. Varsha R Embar , Surbhi despande" sArAmsha - a Kannada Abstractive Summarizer" 2013 International Conference on Advances in Computing, Communications and Informatics (ICACCI)
4. 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.
5. Gupta, Vishal, and Gurpreet Singh Lehal. "Automatic Text Summarization System for Punjabi Language." Journal of Emerging Technologies in Web Intelligence, pp. 257-271, 2013.
6. 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.
7. 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.
8. Dhanya, P. M., and M. Jathavedan, "Comparative Study of Text Summarization in Indian Languages." International Journal of Computer Applications vol.75, no.6,2013.
9. 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.
10. 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.
11. 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.
12. Lloret E., "Text summarization: an overview" Paper supported by the Spanish Government under the project TEXT-MESS (TIN2006-15265-C06-01). 2008.
13. Hans Peter Luhn. "The automatic creation of literature abstracts," IBM Journal of research and development, 2(2):159–165, 1958.
14. Po Hu, Tingting He, and Donghong Ji. "Chinese text summarization based on thematic area detection." In ACL-04 Workshop: Text

BIOGRAPHY

Nilofar Mulani is student of Computer Department, SES'S FOE College of Engineering , Diksal ,Dist. Raigad University of Mumbai Maharashtra ,India and **Sheetal K Dhamal** is Asst.Professor at Lokmanya Tilak College of Engineering,Koparkhairne, Navi Mumbai, Maharashtra, India. Their research interests are Natural Language Processing, Automatic Text Summarization techniques etc.