International Research Journal of Engineering and Management Studies (IRJEMS)
Volume: 03 Issue: 04 | April -2019 ISSN: 1847-9790 || p-ISSN: 2395-0126

Automatic Summarization of Hindi Text Document

Apurve Singh and J. Briskilal

Apurve Singh, Department of Computer Science and Engineering, Kattankulathur Campus, SRM Institute of Science and Technology

apurve.7@gmail.com

Ms. Briskilal.J, Assistant Professor Department of Computer Science and Engineering, Kattankulathur Campus, SRM Institute of Science and Technology
briskilal.j@ktr.srmuniv.ac.in

Abstract -*A Summarization of a document is a miniature version of the document which conveys the essence of the source document. A summary of a document always contains only the important information from the document. This paper targets towards a decent summarization of the given text document, by extracting the most relevant sentences. The sentences are indexed in accordance to take Hindi document as input. It also sheds idea on the detection and removal of Deadwood, and summarization method designed for Hindi text Document based on history information. The algorithm contains lexical association, sentences indexing as well as word indexing.*

***Key Words*: Hindi Text Document, Bernoulli model of Randomness, Context based word Indexing**

1.INTRODUCTION *(Size 11 , cambria font)*

As the Internet is ever-so expanding, the amount of data is also increasing at a huge rate. The problem it faces right now is this sudden increase in information and hence the need of research and advances in automatic text summarization. Instead of reading the whole text document, which consists of example, comparison, support detail, etc., it is always convenient for the readers to read point to point and get the basing gist of it. Automatic Text Summarization is meant for that exact same thing. It summarizes the text into a filtered description of the original text. The summary would be a non-redundant look alike of the original containing the same essence. Summarization can be of two types: Extractive and Abstractive. In the proposed system, Extractive Summarization has been chosen for summarization. The text would be graded or scored upon 'Features'. Features are characteristics of a sentence that it should possess to be included in the summary. In this paper, we have focused more on the pre-processing stage to reduce the summarization much lower.
The Hindi language has been taken as language of study for this paper. Hindi, written in the Devanagari script has one of the largest set of alphabets, is one of the most spoken languages in India. It is the native language of people populating the central part of India, such as Delhi, Chhattisgarh, Himachal Pradesh, Chandigarh, Bihar, Jharkhand, Madhya Pradesh, Haryana, and Rajasthan. During related search, it was shed light that work done on English language for text summarization has lot of resource and work done on it, while Hindi language had too few to count. This motivated me to take Hindi Language as the language of Study

II. LITERATURE SURVEY

1.) Text Summarization for Bengali Language

Islam and Masum developed a text summarization system 'Bhasa' for summarizing Bengali text. A corpus-oriented Summarization system, it is based corpus scores, scored upon words (queries) having the highest frequency, and then creating the summary of the text document on the basis of the words (queries) by applying vector-space-term-weighting. A tokenizer, which determines tags, abbreviation, headings, terms, title etc; tokenizes the original document. Then scoring the tokenized text document is done. Afterwards summarization is done according to the ranking of the tokenized document creating a summary.
Das and Bandyopadhyay developed a Bengali opinion text summarizer. It basically determines the information on the essence of the original text document, then utilizes the same for summarization.Its based on a model that utilizes topic sentiment. This means that the selection of sentences is done by determining and aggregation of the sentiment. Aggregation is achieved by using k-means approach and applying theme graph representation. Both of these are ultimately used for selection of relevant sentence for summarization.

2.) Text Summarization for Punjabi Language

V. Gupta proposed a Punjabi text summarizer that targets the regional language oriented features. In this

summarizer each line on the original text document is treated as a vector of different features, with the features ranging from sentence relative length, Term-frequency, Punjabi terms related to common nouns, Punjabi named locations names, entities, existence of numerical data, etc. Weight of each sentence is then calculated by applying regression, a weight learning method. For each sentence, the score of the feature present in the sentence is calculated and a final score is presented. At last, sentences with top scores are selected for summarization. The borderline weight of allowed sentences is set according to the Compression Ratio allowed by the user.

3.) Text Summarization for Kannada Language

R. Jayashree and, K. M. Srikanta and K. Sunny proposed a Text summarization system for Kannada text. Named "Kannada text Summarizer based on Key terms Extraction", this system utilizes Kannada text documents, already tagged and marked, from web resources and identifies the important and emphatic phrases and sentences. It also uses Inverse-Document-Frequency techniques with Term-Frequency for application on the phrases/sentences for constructing a summary.

R. Jayashree proposed another summarization for Kannada text which has its base in the previous technique, utilizing the Inverse-Document-Frequency techniques with Term-Frequency and then applies them for summarizing the text. It adds giving a final weight to every line and score them for further summarization. Improvement in this summarization is the addition of a database for reviewing Kannada text documents. The Kannada text documents are taken from Webdunia, a web portal in Kannada that contains News of all kind of genre. Summary so constructed will be based in the number of sentences of the original text document. It also removes sentences that means the same.

4.) Text Summarization for Tamil Language

M. Banu, C. Karthika, P. Sudarmani and T.V. Geetha proposed a Tamil text summarization technique. It utilizes the approach of creating a subgraph to create a text summary of the document.The subgraph is utilized for the selection of the sentences to be added in the summarization. The syntax of the language is used for summarizing the text document. Triples of Subject-Object-Predicate are selected from individual lines to from a semantic graph of the original text document. Normalization is used for reducing the frequency of nodes of the semantic graph of the Original Document. The learning technique is based on support vector machine learning

5.) Text Summarization for Hindi Language

U. Garain, A. K. Datta, U Bhattacharya and S.K. Parui proposed Text Summarization of Hindi Text document. This summarization utilizes JBIG2 coded text pictures, without optical character recognition. The images are unwinded and then the whole document is marked sentence by sentence. Four major feature are scored accordingly for the sentences. These features are: [1] Length of Sentence.[2] Position of Sentence [3] Thematic term feature [4] Title term. The presence of these features in a sentence mark them as summary lines or non-summary lines. The summary sentences are grouped together and are further ranked. These 'summary sentences' are then combined to create the summary.

6.) Text Summarization for Assamese Language

Navanath Saharia, Utpal Sharma and Jugal Kalita, adopted suffix stripping approach.
This approach utilizes an engine, based on rules, to generate all probable suffix sequence
Two way to tackle it has been told in the paper:- the First one completely based on the rule engine, but has a problem of over-summarization or under-summarization with this approach. The Second one utilizes a frequent root-word with suffix stripping to give a refined outcome.

7.) Text Summarization for Gujarati Language

The Author has discussed the preprocessing phase of Text Summarization of Gujarati Text. The approach is a rigid stepwise process and targets primarily towards the removal of certain words as well as recognition and removal of duplicate sentences. They progress by removing hyphen, stop word, duplicate word. the last step is the recognition and removal of duplicate sentences.

8.) Text Summarization for Odia Language

The author has used extraction method for this Odia Text summarization. Extraction is done on the weight distribution of the sentences. The only catch is to take the first sentences of the paragraph as it would be the most relevant to the topic, and will contain a proper noun. Hence they give the first sentence the maximum weight.

9.) Text Summarization for Marathi Language

This paper proposes summarization of newspaper articles for summarizing. An algorithm, The Keyword Extraction Algorithm works to find the most used or top scored words efficiently and using same data summarizes the article. The size of the summarized article is based on the length of the original article.

III. PROPOSED SYSTEM

The proposed system would be a rule or grammar-based approach. Some rules would be made to generate the summary of the Hindi text document. A group of tables (corpus) for Hindi Language will also be used along with the rules to extract the important lines from the document. The corpus would contain the following set of tables: -

1. People Names.
2. Names of Places.
3. City names
4. State and Country names
5. Dead Phrases.

The basic rule to approach Hindi text Document would be as follows: -

I. Input the Hindi text document
II. Divide the whole text into paragraphs and then into lines with the help of the punctuation used.
III. Remove the dead phrase from the document and replace them.
IV. For each line calculate equivalent weight of that line using algorithm and the corpus.
V. Remove lines from the document whose weight is less than the required minimum weight.
VI. Combine the remaining lines to form a summary.

In general, the sentences with less weight, in essence the sentence with least regards to the document would be removed and not included in the summary. Also, the dead phrases so present in the document is also replaced with words stored in the corpus and hence are not included in the summary of the Hindi text document.

REFERENCES

1. *T. Islam and S. M. A. Masum, "Bhasa: A Corpus Based Information Retrieval and Summarizer for Bengali Text," Macquarie University, Sydney, Australia, 2004.*
2. *Das and S. Bandyopadhyay, "Topic-Based Bengali Opinion Summarization", International Conference COILING '10, Beijing, pp. 232–240, 2010.*
3. *V. Gupta and G.S. Lehal, "Automatic Text Summarization for Punjabi Language," International Journal of Emerging Technologies in Web Intelligence, vol. 5, pp. 257-271, 2013.*
4. *V. Gupta and G. S. Lehal, "Complete Preprocessing Phase of Punjabi Language Text Summarization," International Conference on Computational Linguistics COLING'12, IIT Bombay, India, pp. 199-205, 2012.*
5. *V. Gupta and G. S. Lehal, "Automatic Punjabi Text Extractive Summarization System," International Conference on Computational Linguistics COLING '12, IIT Bombay, India, pp. 191-198, 2012.*
6. *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.*
7. *R. Jayashree, "Categorized Text Document Summarization in the Kannada Language by Sentence Ranking," Proceedings of 12th International Conference on Intelligent Systems Design and Applications (ISDA), pp. 776-781, 2012.*
8. *M. Banu, C. Karthika, P. Sudarmani and T.V. Geetha, "Tamil Document Summarization Using Semantic Graph Method", International Conference on Computational Intelligence and Multimedia Applications, IEEE, pp. 128- 134, 2007.*
9. *U. Garain, A. K. Datta, U Bhattacharya and S.K. Parui, "Summarization of JBIG2 Compressed Indian Textual Images," Proceeding of 18th International Conference on Pattern Recognition (ICPR'06), IEEE, Kolkata, India, 2006.*
10. *Pawan Goyal, Laxmidhar Behera, Senior Member, IEEE, and T. M. McGinnity, Senior Member, IEEE "A Context based Word Indexing Model for Document Summarization"*
11. *Navanath Saharia, Utpal Sharma and Jugal Kalita "Analysis and Evaluation of Stemming algorithms: a case study with assamese". icacci'12, august 3-5, 2012, chennai, t nadu, india.*
12. *Ashish B. Tikarya, Kothari Mayur, Pinkeh H. Patel "Pre-Processing Phase of Text Summarization Based on Gujarati Language" International Journal of Innovative Research in Computer Science & Technology (IJIRCST) ISSN: 2347-5552, Volume-2, Issue-4, July - 2014*
13. *R. C. Balabantaray, B. Sahoo, D. K. Sahoo, M. Swain "Odia Text Summarization using Stemmer" International Journal of Applied Information Systems (IJAIS) – ISSN : 2249-0868 Foundation of Computer Science FCS, New York, USA Volume 1– No.3, February 2012*
14. *Shubham Bhosale, Diksha Joshi, Vrushali Bhise, Rushali A. Deshmukh "Marathi e newspaper text summarization using automatic keyword extraction technique" International Journal of Advance Engineering and Research Development (IJAERD) Volume 5- No.3, March 2018*