Marathi Text summarizer using Text rank algorithm

Abstract

Manual summarization of huge documents of texts is tedious and error-prone. Also, the results in such quite summarization may cause different results for a selected document. Thus, Automatic text summarization has become important because of the tremendous growth of knowledge and data. It chooses the foremost informative neighbourhood of text and forms summaries that reveal the foremost purpose of the given document. It yields a summary produced by a summarization system which allows readers to understand the content of the document instead of reading each and every individual document. The aim of Text Summarizer is to supply the meaning of the text in fewer words and sentences. Summarization is often categorized into Abstractive and Extractive. This project is based on an extractive concept implemented on the studied models.

Numerous text summarization systems are there today for English or other languages. But when it involves Indian languages, we observe an inadequate number of automatic summarizers. Our efforts during this direction are mainly for developing an automatic text summarizer for the Marathi Language. We glance forward to gauge the obtained summaries using the ROUGE metric.

Keywords: summarizer, extractive, Marathi, text rank, features extraction.

Introduction

Text Summarization could also be a way of condensing actual text into an abstract form that provides the same meaning and knowledge as provided by the actual text. It chooses the foremost informative neighbourhood of text and forms summaries that reveal the foremost purpose of the given document. It yields a summary produced by a summarization system which allows readers to understand the content of the document instead of reading each and every individual document. There are 2 types of summarizer: Extractive and Abstractive summarizer.

Extractive text summaries focus on extracting appropriate sentences from the source text in a sequential manner. The acceptable sentences are extracted by applying statistical and language reliable features to the input text. But there's the limit in extraction. The extracted phrases and sentences are set in chronological order.

Abstractive text summaries are formed by enacting tongue understanding concepts. This type of summarizer generally, incorporates terms that don't exist within the document. It aims to imitate methods employed by humans, like representing an idea that's available within the original article in a better and more comprehensive way. It’s an effective summarizer however, it's very difficult to implement.so in this paper we will focus on extractive based summarizer.

Motivation

Various automatic text summarization systems are accessible for several often used languages. Many of the summarizers are for other foreign languages. Moreover, technical documentation is usually minimal or may be absent. When it involves Indian languages, automatic summarization systems are limited. Little or no research and work has been exhausted text summarization for the Indian language Marathi (an Under-Resourced language).

Problem statement

It’s always hard to summarize huge documents manually and in this era of automation we need something which will help us reduce this work .The approach proposed in this paper is to build extractive summarizer where we will comprehend the content of document by extracting crucial sentences or passages from the document and with the use of text rank algorithm.

Literature Survey-

- Virat V. Giri, Dr. M. M. Math, and et al., “Indian language based summarization a method proposed by them also studied the phases involved in the process of converting any document into its summary. In their study, they proposed a method for Marathi summarizer in detail. They also proposed processes, pre-process on that document including Marathi Stemmer, Proper name and noun list in Marathi, also Marathi keywords extraction, rule-based entity recognition, and many others. Pre-processing phase followed by processing phase. Process including weight analysis of words depending on features by regression analysis.’’

- Hamzah Noori Fejer , Nazlia Omar and et al,” Extraction of key phrases and technique of clustering this new approach is proposed. . Their new approach of clustering which combines hierarchical and k-means clustering. The results obtained from their experiments are their proposed model gives better performance when compared with existing approaches”.

- Deepali K. Gaikwad, and et al.," This paper introduced technique which gives a summary of any text using rule-based stemmer technique or generating question, they used rule-based approach of abstractive text summarization. They used rule-based stemmer as well as POS tagger and NER tools. In this approach Marathi text is taken as input, POS tagger is applied on it and then Marathi language rule-based `who` type questions are generated for the given input.

- Ms. Jayshri Arjun Patil and et al., “Various named Entity Recognition data and also gives detailed information regarding challenges and problem arising with Indian languages. The Proposed approach towards discovering a named entity in a document and then categorize these NEs into diverse Named Entity classes like Name of Person, Location, River, Organization, etc. They give a brief introduction to Name Entity Recognition”.

- Rafael Ferreira, Luciano de Souza Cabral and et al.’’ here they compared more than 15 algorithm’s qualitative and quantitative assessment, which are used for sentence scoring available in the literature i.e. News, Blogs and Article contexts. Also gives directions to improve the sentence extraction results obtained are suggested. They used a four-dimensional graph-based model for text summarization which relies on four dimensions (similarity, semantic similarity, coreference, discourse information) to create the graph”.

System architecture

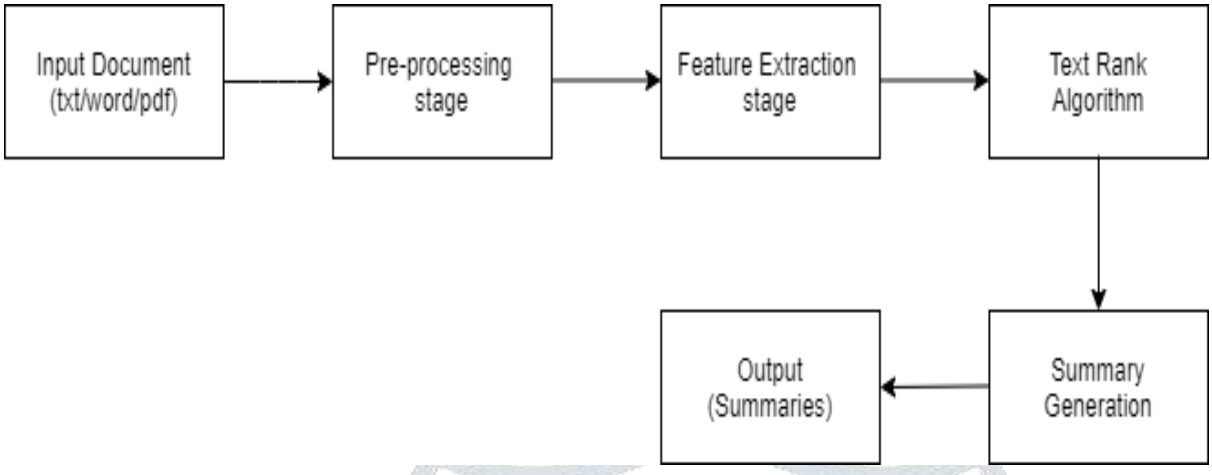

Figure 1 : System Architecture

Module

1. Dataset: The EMILLE (Enabling Minority Language Engineering) which includes monolingual, parallel and annotated corpora for Asian Languages including Marathi will be used as dataset for extracting summaries.

2. Pre-processing Stage: Pre-processing stage is a very important stage in Marathi text summarization. In this stage we obtain pre-processed data which will be best suited for our model. In general the Pre-processing stage consists of four steps which include removal of punctuation marks, stop word removal, stemming and tokenization.

3. Feature extraction: In this step important features like sentence positional value (POS tag), SOV (Subject Object Verb), TF-IDF (Term Frequency/ Inverse Document Frequency) or TF-ISF (Term Frequency/ Inverse Sentence Frequency) are extracted from Pre-processed data. These features will be used to rank sentences based on their features extracted.

4. Scoring: Sentence Scoring is the process of assigning scores to each of the sentences from the article, denoting its importance from the given article. The score is calculated based upon features extracted.

   This phase involves:

   1) Identifying the feature given by the user and extracting sentences accordingly.

   2) Assign the score to the extracted sentence by using sentence scoring technique.

$$\text{Score Sentence} = \sum_{i=1}^{n} \text{Score Features}$$

5. Graph Scoring: In this method, score is calculated on the basis of the relationship found between the sentences. The given model includes the Text Rank algorithm for the summarization of the article. The steps are as follows

1. Obtain the text units that best define the summary we need, and add those units as vertices in the graph.
2. Find relations between sentences that connect such text units to draw edges as connectors between those vertices.
3. Iterate the graph based text ranking algorithm until all possible connections are achieved.
4. Classifying the vertices based on to their final obtained score. Utilize the values attached to each vertex for ranking/selection purposes.
5. Create a summary based on top ranked sentences.

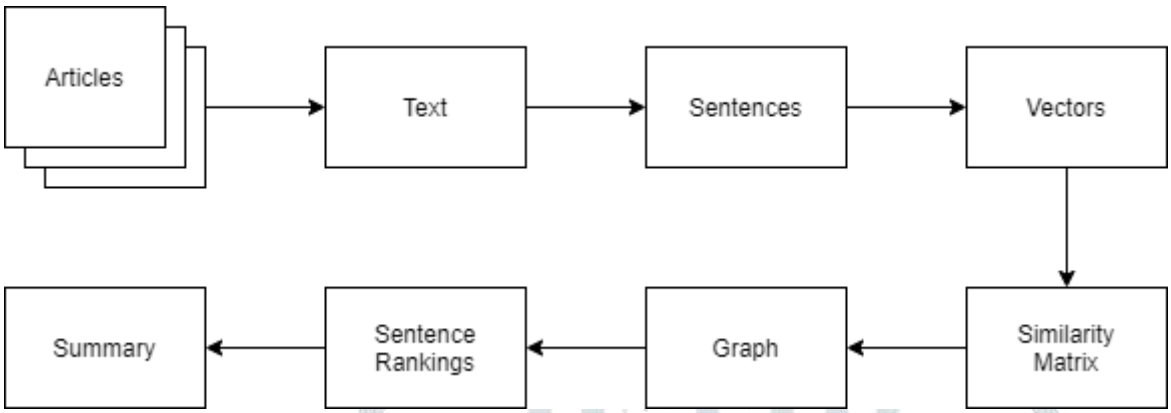

Figure 2: Working of Text Rank algorithm

Algorithm

Text Rank Algorithm:

Text Rank is an unsupervised extractive summarization technique which works similar to the page rank algorithm. It is a graph based algorithm in which sentences are the vertices of a graph and the edges act as relations between the sentences.

Working of text rank algorithm is as follow

1. Concatenate all the text contained in the article which is required for Summarization.
2. Split the text from the article into individual sentences.
3. Defining vector representation of each and every statement of the article.
4. Store the similarities found between the sentences in the matrix.
5. Similarity matrix then converted into the graph.(sentences act as a node while relations act as edges between them).
6. Make a summary from top-ranked sentences.

Conclusion

With the tremendous increase in the amount of content accessible online, there is a need of fast and effective automatic summarization system. The most important steps in extractive summarizations are feature extraction, scoring and clustering. This system can be used in various fields like education, in search engines to improve their performances, for Marathi news clustering.

References

- [1] Virat V. Giri, Dr.M.M. Math and Dr. U. P. Kulkarni , " A Survey of Automatic Text Summarization System for Different Regional Languagesin India ", Bonfring International Journal of Software Engineering and Soft Computing, Vol. 6, Special Issue, October 2016.
- [2] Hamzah Noori Fejer and Nazlia Omar, "Automatic Multi-Document Arabic Text Summarization Using Clustering and Key phrase Extraction", ICIMU IEEE 2014 International Conference.
- [3] Deepali K. Gaikwad, Deepali Sawane and C. Namrata Mahender, " Rule Based Question Generation for Marathi Text Summarization using Rule Based Stemmer ",IOSR Journal of Computer Engineering (IOSR-JCE), 2015.
- [4]Ms. Jayshri Arjun Patil, Ms. Poonam Bhagwandas Godhwani, "Review of Name Entity Recognition in Marathi Language", IJSART - Volume 2 Issue 6, June 2016.
- [5]Rafael Ferreira, Luciano de Souza Cabral, "Assessing sentence scoring techniques for extractive text summarization Expert Systems with Applications", Elsevier 2013.
- [6] Sunitha C, Dr. A Jaya and Amal Ganesh, "A Survey of Abstractive Summarization Techniques in Indian Languages", 2016.