International Organization of Research & Development (IORD)
ISSN: 2348-0831
Vol. 09 Issue 01 | 2021

Marathi Text Document Summarization using Neural Networks

[1]Prof. Prashant Itankar, [2]Prof Ankit Sanghavi, [3]Ms. Anushree Mane

[1,2,3] Alamuri Ratnamala Institute of Engineering & Technology, Shahapur, Thane

DOI: 10.5281/zenodo.5587744

ABSTRACT

The internet is comprised of web pages, news articles, status updates, blogs and much more. It is difficult to navigate through this data as it is unstructured and usually discursive. Manual summarization of large documents of texts is tedious and error prone. Also, the results in such kind of summarization may lead to different results for a particular document. Thus, Automatic text summarization has become important due to the tremendous growth of information and data. It chooses the most informative part of text and forms summaries that reveal the main purpose of the given document. It yields summary produced by summarization system which allows readers to comprehend the content of document instead for reading each and every individual document. So, the overall intention of Text Summarizer is to provide the meaning of text in less words and sentences. To perform extractive text summarization we propose to use a Recurrent Neural Network (RNN) – a type of neural network that can perform calculations on sequential data (e.g. sequences of words) – as it has become the standard approach for many Natural Language Processing tasks.

***Index Terms*— Text summarizer, Neural machine translation, NLP.**

1. INTRODUCTION

Unlike existing solutions, we build this work on the The main use of text summarization is to help readers save time and effort of reading long documents to find useful information. Text summarization is creating a short, accurate and consistent summary of a longer document. We cannot create summaries of all the text manually and there is a great demand for automatic method. Summarization helps users in several ways such as reducing reading time, making selection procedures faster and easier for researching documents, improves effectiveness of indexing. Automatic summaries are less biased as compared to human summaries. Summarization programs and systems are commercially in demand by the military, universities, research institutes, law firms, etc.

Text Summarization is a technique of condensing actual text into abstract form which provides same meaning and information as provided by actual text. It chooses the most informative part of text and forms summaries that reveal the main purpose of the given document. It yields summary produced by summarization system which allows readers to comprehend the content of document instead for reading each and every individual document. So, the overall intention of text summarizer is to provide the meaning of text in less words and sentences. Summarization systems can be sorted into two categories: Abstraction-based summarization and Extraction-based summarization. Extractive summaries involve extracting appropriate sentences from the source text in sequential manner. The appropriate sentences are extracted by applying statistical and language reliable features to the input text. But there is limit in extraction. The extracted phrases and sentences are in chronological order. While, abstractive text summaries are formed by enacting natural language understanding concepts.

This kind of summarizer generally, incorporates terms that do not exist in the document. It aims to imitate methods used by humans, such as representing a concept that is available in the original article in a better and more comprehensive way. It is effective summarizer however, it is very difficult to implement.

2. LITERTURE SURVEY

Extractive Text Summarization is the method of extracting content from the document and combining it to form a text smaller in size. This ensures that only the words having relevance in the document are selected for the summarization. Abstractive Text Summarization takes summarization to a stride further. It is capable of depicting information by creating new sentences. Abstractive Summarization can be divided into Structured and Semantic approaches.

Each of these classifications can be subdivided into subcategories based on various methods. In tree-based approaches, clustering from many documents occurs. This clustering is done with the help of the order and significance of these documents. Linearization is used for the formation of sentences using tree traversal [14]. Ontology-based approach is used for creating domain-related summaries. A domain or a dataset is fed to the

system in advance, and based on this dataset, the system generates summaries with relevant text in the summary [3].

Rule-based method is based on random forest classification and feature scoring. The scoring is based on the constraints laid down by the user. The rules can be set in many ways, such as: using verbs and nouns which are related to each other; keywords and syntactic constraints; domain constraints [7]. Graph-Based Approach uses the graph data structure for language representation. Here, every word unit is represented by a node, and the structure of the sentences is determined by directed edges. These edges represent the relationship between any two words. The underlying feature of this method is that it uses the shortest path algorithm to find the smallest sentences with a considerable amount of information. The sentence formation is subjected to constraints such as it is mandatory to have a subject, verb, and predicate in it. Along with this, a compendium is used for Linguistic and Summary Generation purposes [1].

Information Item based approach is said to be the smallest element of coherent information in a sentence. Text entries, its attributes, and, predicates are identified in this method. Similar to Extractive Text Summarization Methods, Frequency-based models are used for item set selection. The sentences are generated by combining them, and ranking of sentences is done. Out of these sentences, the highly ranked content is selected to be a part of the summary[17].

Semantic Text Representation Model is based on aims to analyze input text using the semantics of words rather than the structure of the text. Here, the abstractive summarization is accomplished as a semantic portrayal of supply records.Content determination is finished by the selection of the most relevant predicate contention structures. At last, the summary is created by utilizing a dialect apparatus. However, the framework does not deal with comprehensive semantics in the summarization technique [9].

Abstractive Summarization has been achieved using a sequence to sequence encoder-decoder model. This model has its famous application in Neural Machine Translation. These language models are capable of taking an input of size N and give an output of size M. It is primarily used to preserve the dependencies LSTM cells are used. These cells are the most atomic unit of an encoder and decoder. Similar to LSTM, GRU cells can also be used at the expense of some accuracy.

There are ways in which Encoders have been designed over a few years. The most basic implementation is to use a bag of words. These encoders are capable of capturing crucial words while ignoring the relationship between the neighboring words.Attention Mechanism can be implemented in two ways wherein either the same hidden units can be used for computing the attention weights, or certain hidden units can be kept aside specifically for calculation of Attention Weights.

3. RELAED WORK

Various automatic text summarization systems are accessible for most often used languages. Most of these text summarization systems are for English and other foreign languages. Moreover, technical documentation is often minimal or even absent. When it comes to Indian languages, automatic summarization systems are very limited. Very little research and work has been done in text summarization for the Indian language marathi (an Under-Resourced language). To perform and depend on the answers that are provided by the student, it will generate a summarized result. The main aim of this system is to provide an overview of the Artificial Intelligence techniques that we used to predict the performance of the student.

4. PROPOSED APPROACH

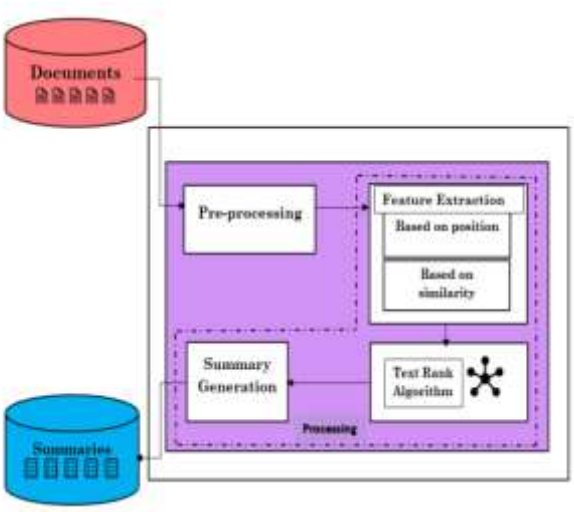


Fig. 1 Block Digram

4.1 Modules

- Document Preprocessing.
- Feature Extraction.
- Sentence Scoring.
- Graph Scoring.
- Similarity
- Summarization.

5. METHODOLOGY

Text Summarization is a technique of condensing actual text into abstract form which provides same meaning and information as provided by actual text. It chooses the most informative part of text and forms summaries that reveal the main purpose of the given document. It yields summary produced by summarization system which allows readers to comprehend the content of document instead for reading each and every individual document. So, the overall intention of text summarizer is to provide the meaning of text in less words and sentences. Summarization systems can be sortedinto two categories: Abstraction-based summarization and Extraction-based summarization.

5.1 Dataset

Multiple documents from dataset will be used to extract texts from documents on different subjects, such as education, politics and news articles. We have 634 documents based on news articles discussing the different topic in Marathi language.

The EMILLE (Enabling Minority Language Engineering) which includes monolingual, parallel and annotated corpora for Asian Languages including marathi is used for obtaining multi documents. The system can be divided into two broad stages: Pre-Processing and Processing stage.

5.2 Pre-processing Stage

Pre-processing stage is essential in text summarization.It results into pre-processed data, which is ideally fit for processing stage. In general pre-processing stage consists of steps to remove punctuation marks, tokenization, stop word removal, stemming, etc. In this section we will discuss various steps used in pre-processing stage.

1) Boundary identification and punctuation marks removal:

Every sentence ends with a punctuation mark depending on the nature of sentence, whether interrogative, exclamatory, imperative or declarative. Also, use of quotation marks (" ",'), commas(,), special characters(&,*,—) and symbols(#,@), etc.is frequent. But when it comes to extract important words for processing stage, we need to eliminate these punctuation marks. Hence, we use techniques for removal of punctuation marks. The output of this step is punctuation marks free sentences in the document.

2) Stop words elimination: Frequently occurring non essential words for processing in text summarization are generally termed as stop words. In marathi language, we use stop words like shivay, ase, eetar etc. in day to day use. We should eliminate them for obtaining meaningful context while processing the documents. The output of this sentence is stop words free sentences in the document.

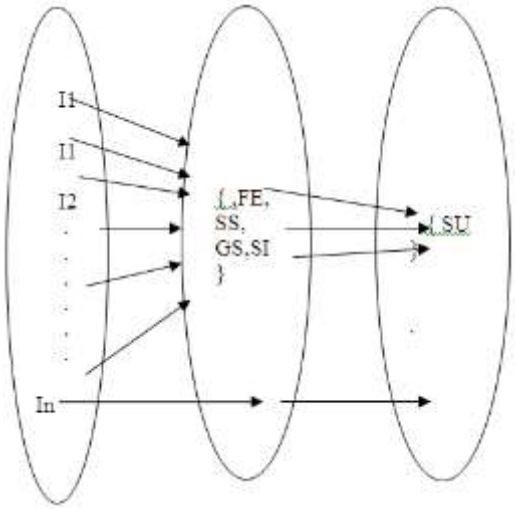
3) Stemming and lemmatization: The process of obtaining stem / radix or root word for morphological variants present in the documents. Lemmatization identifies lemma of a word. It is mapping of verbs into their infinitive and nouns into their singular form. Methods used for constructing stemmers include : Rule based-Porter's Stemmer, Husk stemmer, Unsupervised stemming, suffix stripping-Lovins stemmer, Dawson stemmer, N gram method, HMM method, YASS(Yet Another Suffix Stripper) stemmer etc.

5.3 Feature extraction

The features like SOV (Subject Object Verb - Experimental) verification, sentence positional value (POS tagging), TF-ISF (Term Frequency/ Inverse Sentence Frequency) or TF-IDF (Term Frequency/ Inverse Document Frequency) are extracted from pre-processed sentences. Sentences are further ranked on basis of features extracted.

6. MATHEMATICAL MODEL

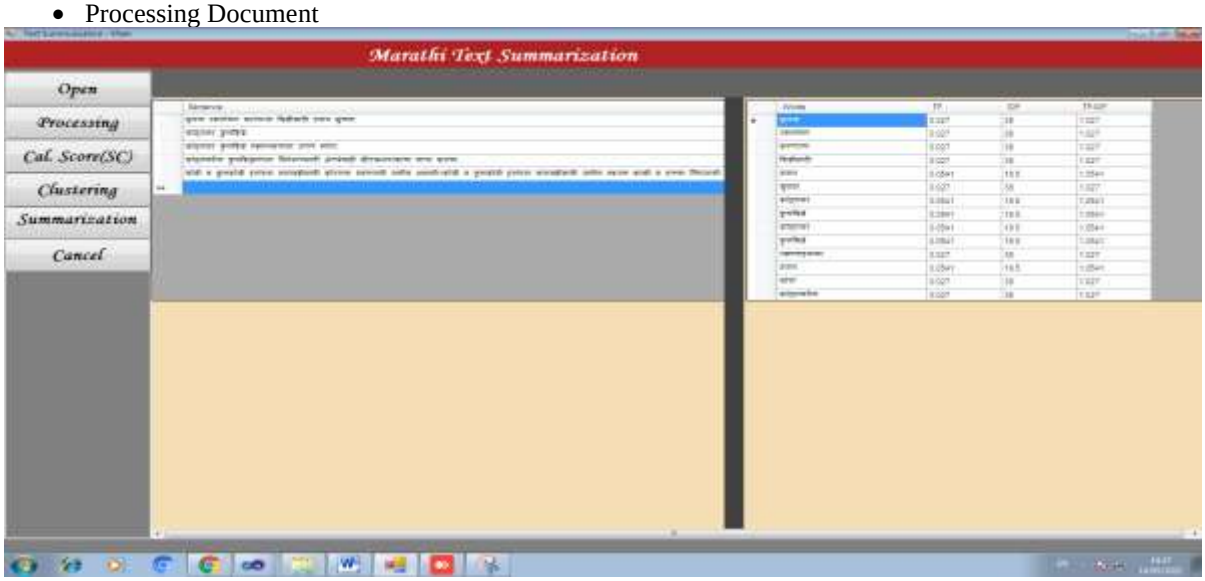
S: is a System.
DI: Document Input.
FE: Feature Extraction.
SS: Sentence Scoring.
GS: Graph Scoring.
SI: Similarity.
SU: Summarization.
S={DI,FE, SS, GS,SI, SU}



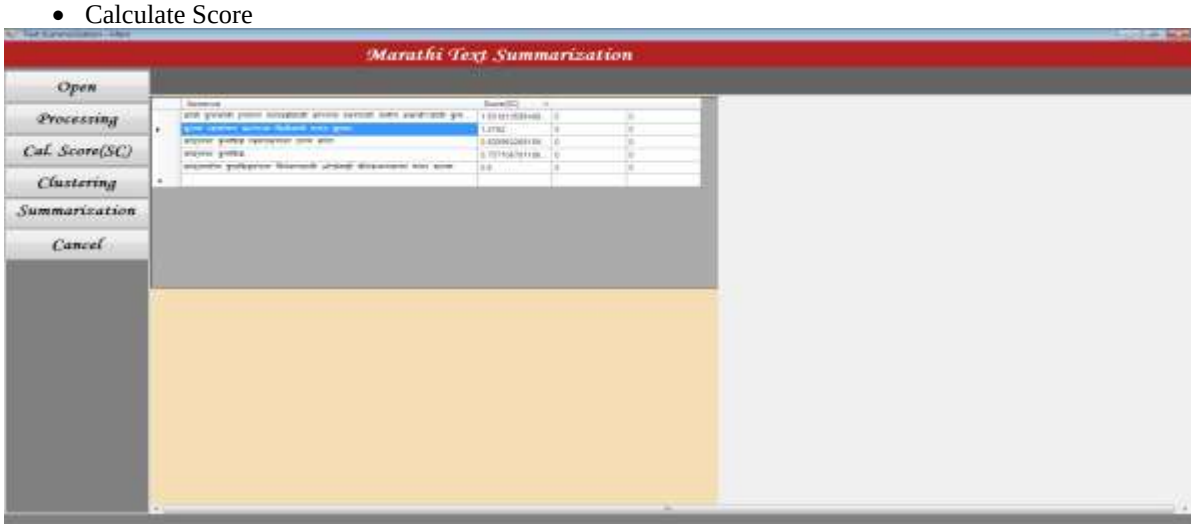
7. RESULTS

- File Read

- Processing Document



- Calculate Score



- Clustering

- Summarization

8. 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 this system approach are feature extraction, scoring and graph generation. This system can be used in various fields like education, in search engines to improve their performances, for Marathi news clustering, Question generation purpose and many other application oriented areas, etc.

9. REFERENCES

[1] Ganesan, K., Zhai, C., & Han, J. (2010, August). "Opinosis: a graph-based approach to abstractive summarization of highly redundant opinions." In Proceedings of the 23rd international conference on computational linguistics (pp. 340-348). Association for Computational Linguistics.
[2] Rahimi, Shohreh Rad, Ali Toofanzadeh Mozhdehi, and Mohamad Abdolahi. "An overview on extractive text summarization." Knowledge-Based Engineering and Innovation (KBEI), 2017 IEEE 4th International Conference on. IEEE, 2017.
[3] Lee, C. S., Jian, Z. W., and Huang, L. K. (2005). "A fuzzy ontology and its application to news summarization." IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 35(5), 859-880
[4] Andhale, Narendra, and L. A. Bewoor. "An overview of text summarization techniques." Computing Communication Control and automation (ICCUBEA), 2016 International Conference on. IEEE, 2016.
[5] Barrios, Federico, et al. "Variations of the similarity function of textrank for automated summarization." arXiv preprint arXiv:1602.03606 (2016).
[6] Nallapati, Ramesh, Bowen Zhou, and Mingbo Ma. "Classify or select: Neural architectures for extractive document summarization." arXiv preprint arXiv:1611.04244 (2016).
[7] John, A., and Wilscy, M. (2013, December). Random forest classifier based multi-document summarization system. In Intelligent Computational Systems (RAICS), 2013 IEEE Recent Advances in (pp. 31-36). IEEE.
[8] Jain, Aditya, Divij Bhatia, and Manish K. Thakur. "Extractive Text Summarization Using Word Vector Embedding." Machine Learning and Data Science (MLDS), 2017 International Conference on. IEEE, 2017.
[9] Khan, Atif, Naomie Salim, and Yogan Jaya Kumar. "A framework for multi-document abstractive summarization based on semantic role labelling." Applied Soft Computing 30 (2015): 737-747.
[10] Moratanch, N., and S. Chitrakala. "A survey on abstractive text summarization." Circuit, Power and Computing Technologies (ICCPCT), 2016 International Conference on. IEEE,2016.
[11] Zhang, P. Y., and Li, C. H. (2009, August). "Automatic text summarization based on sentences clustering and extraction." In Computer Science and Information Technology, 2009. ICCSIT 2009. 2nd IEEE International Conference on (pp. 167-170). IEEE.
[12] Narayan, S., Papasarantopoulos, N., Lapata, M., and Cohen, S. B. (2017). "Neural Extractive Summarization with Side Information." arXiv preprint arXiv:1704.04530.
[13] Gupta, V., and Lehal, G. S. (2010). "A survey of text summarization extractive techniques. Journal of emerging technologies in web intelligence", 2(3), 258-268.
[14] Barzilay, R., and McKeown, K. R. (2005). "Sentence fusion for multidocument news summarization. Computational Linguistics", 31(3), 297-328.
[15] Nallapati, Ramesh, et al. "Abstractive text summarization using sequence-to-sequence rnns and beyond." arXiv preprint arXiv:1602.06023 (2016).
[16] Paulus, Romain, CaimingXiong, and Richard Socher. "A deep reinforced model for abstractive summarization." arXiv preprint arXiv:1705.04304 (2017).
[17] Genest, P. E., and Lapalme, G. (2011, June). "Framework for abstractive summarization using text-to-text generation." In Proceedings of the Workshop on Monolingual Text-ToText Generation (pp. 64-73). Association for Computational
Linguistics.
[18] Peddinti, V., Povey, D. and Khudanpur, S., 2015. "A time delay neural network architecture for efficient modeling of long temporal contexts." In Sixteenth Annual Conference of the International Speech Communication Association.
[19] Cheng, J. and Lapata, M., 2016. "Neural summarization by extracting sentences and words." arXiv preprint arXiv:1603.07252