MARATHI TEXT SUMMARIZATION USING MACHINE LEARNING

[1] Utkarsh Hajare Student, Computer Engineering Dept., MESCOE, Maharashtra, India
[2] Devendra Bangade Student, Computer Engineering, Dept., MESCOE, Maharashtra, India
[3] Sanket Rajgiri Student, Computer Engineering, Dept., MESCOE, Maharashtra, India
[4] Prakash Dongare Student, Computer Engineering, Dept., MESCOE, Maharashtra, India
[5] Prof. Shilpa Khedakar, Assistant Professor, Computer Engineering, Dept., MESCOE, Maharashtra, India

ABSTRACT

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. Summarization can be categorized as: Abstractive summarization and Extractive summarization. This case study is based on an extractive concept implemented on the studied models. Numerous automatic text summarization systems are handy today for English and other foreign languages. But when it comes to Indian languages, we observe inadequate number of automatic summarizers. Our efforts in this direction are mainly for developing automatic text summarizer for marathi Language. We look forward to evaluate the obtained summaries using ROUGE metric. This paper describes a multi document marathi extractive summarizer. Keywords- Marathi Text Summarizer, Extractive Summarization, Graph based model, Feature Extraction, TextRank

Keyword : - *Convolutional , Neural, Networks.*

1. Introduction

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.

1.1 Abstractive Summarization:

In this type of summarization, language understanding tools are used to generate a summary. The main focus is on choosing phrases and lexical chains from the documents. General steps used in this technique are withdrawing basic features, obtaining the relevant information, revising and reducing information. Since the formulation of this technique in mathematical or logical form is cumbersome, it is referred as a complex technique to implement . Also, the quality of generated summaries relies on the depth of linguistic strength. These techniques are generally categorized as Structured and Semantic based. Structured based approaches, obtain most significant information from the documents through cognitive schemas such as templates, frames and scripts [3].

Semantic based approach makes use semantic depiction of documents which is further used to supply into natural language generation (NLG) system as input. This method focuses on obtaining noun phrases and verb phrases by managing linguistic data. Phrases thus obtained are then related to concepts, attributes and relations of a domain-specific ontology. The important document areas (like sentences or paragraphs) are selected by using ontology-based annotations and clustering techniques. The information obtained as an outcome is used to transform those areas into semantic representation. This NLG system takes this representation as input and then produces abstracts.

Chart -1: Rouge Result

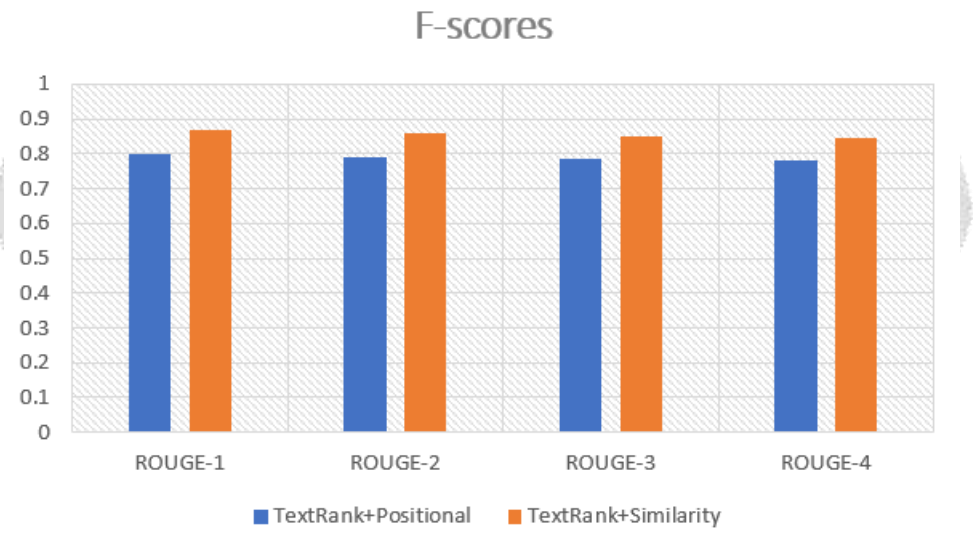

Table -1:ROUGE RESULTS

Technique	ROUGE-1	ROUGE-2	ROUGE-3	ROUGE-4
TextRank+Positional	0.8004	0.7909	0.785	0.7799
TextRank+Similarity	0.8684	0.8602	0.8521	0.8447

Word parsing, rhetorical parsing, statistical parsing and a mixture of all are some of the common techniques used.

The drawbacks of abstractive summarization approach are

1) Machine generated automatic summaries would result into lack of clarity even within a sentence as sentence synthesis is an emerging field.

2) As they heavily depend on the adaptation of internal tools to perform information extraction and languagegeneration, they are difficult to replicate [3].

3) Abstracting is not easier as it needs semantic understanding of content present in the document.

1.2 Extractive Summarization

Extractive summaries are formed by extracting crucial text (sentences or passages) from the document, based on features such as word/phrase frequency, location or cue words to locate the sentences to be extracted. The most important text is treated as the most frequent or the most suitably positioned text. Such an approach thus, shuns the labour on depth of content understanding. They are theoretically simple and easy to implement. This system constitutes two important phases, which are : Pre-Processing and Processing phase

1) Representing the text in a structured manner is the main aim of Pre-Processing phase .

2) Processing phase represents various features that decide the importance of sentences. Certain statistical features used for marathi language are keywords identification, sentence length feature and numerical literals count feature. An equation of summation of feature weights is used to generate score of sentences and high scored sentences in a specific order of input text are considered for final summary. This report describes multi document marathi extractive summarizer. It is text extraction based summarization system which is used to summarize the marathi document by retaining the appropriate sentences based on features.

2. Related Work

A. The development of new techniques is always required to help solving problems [17-31]. Many different techniques have been proposed for automatic text summarization that exploit a variety of different methods. Most of these techniques are extractive text summarization approaches. The modes proposed in [18] use shallow features for text units scoring and choosing text units that have highest scores as summary. Proposed models in [19] use methods to weight scoring coefficient in text unit scoring based on data by machine learning techniques for automatic text summarization. Discourse structure model is proposed in [20] to score sentences.

B. Virat V. Giri and et al. reviewed text summarizers based on various Indian languages and their performances. They studied and proposed summarization method for marathi in detail wherein marathi stemmer, marathi proper name list, English- Marathi noun list, marathi keywords extraction, marathi rule based named entity recognition etc. for pre-processing of text followed by processing of text [1]. Sheetal Shimpikar and et al. studied various techniques of text summarization for various Indian languages [2]. Sunitha C and et al. Worked on Abstractive summarization methods that are used for Indian languages. They explained Abstractive summarization technique, classified in two approaches such as structure based approach and semantic based approach [3]. Hamzah

Noori Fejer and et al. gave a major contribution by proposing a combined approach of clustering technique and extracting keyphrases. They have proposed a new approach of clustering which combines hierarchical and k-means clustering. The results obtained from their experiments proved the proposed model gives better performance when compared with existing ones [4]. An unsupervised approach for marathi stemmer has been discussed by Mudassar Majgaonker and et al. [6]. The present work on text summarization of marathi text with question based system using rule based stemmer technique. For generating question, we used rule based approach of abstractive text summarization and POS tagger, NER tools and rule based stemmer. Here marathi text is taken as input, on it POS tagger is applied and then questions are generated for the given input as per marathi language rules by Deepali K. Gaikwad and et al. At this stage they have framed rules of stemmer only for w´ ho´type questions [5]. Thus it can be extended to learning all What type questions too.

C. Mangesh Dahale proposed text summarizer using inverted indexes [9]. Jayshri Patil and et al. reviewed different approaches of Named Entity Recognition (NER) and discussed issues and challenges arising in Indian languages [8]. Pooja Pandey and et al. discussed extraction of root words using morphological analyzer for devanagari script [11]. Aishwarya Sahani and et al. contributed to automatic text categorization of marathi language documents [7]. Rafael Ferriera et. Al used four dimensional graph based model for text summarization which relies on four dimensions(similarity,semantic similarity,co-reference,discourse information) to create the graph [16]. Federico Barrios et. al used variations in similarity measures along with TextRank for summarization [15]. Our work includes use of TextRank along with positional distribution of sentence scores and considering thematic similarity which gave promising results.

D. Some text summarization techniques are developed base on fuzzy logic method. To get better result and improve the quality of summary, many approaches exploit a combination of two or more different methods [22-24]. In [25] authors propose a model that benefits advantage of diversity based method to pick up the most diverse sentences, and also the advantage of non diversity method which uses fuzzy logic and swarm intelligence to select the most important sentences for text summarization.

E. Since recently many new and powerful machine learning techniques are developed which are mainly based on deep learning methods, some text summarization techniques have proposed to make abstractive summaries that benefits new machine learning models. In [26, 27] authors propose a model which uses an encoder-decoder approach to learn the representation of sentences by encoder and to classify each sentence by decoder based on encoders representations using an attention technique. Proposed model in [28] has two parts. First part is a single sequence model with no decoder to be trained extractively. Second part has a decoder which is abstractively trained to generate sentence-extraction probabilities.

F. Sequence-to-sequence models based on deep learning techniques are used in some abstractive text summarization works. In [29] authors propose a technique which exploits convolutional models to encode the source, and then the abstractive summary will be generated using a context-sensitive attentional feed-forward neural network. In [30] authors develop an abstractive text summarization model based on a sequence-tosequence model and applying the attentional encoderdecoder RNN (Recurrent Neural Network). [31] proposes a strong model that uses a hybrid pointergenerator network to copy words from the source text by pointing, in the first phase, and then exploits coverage to keep track of summary, that prevents repetition, in the second phase.

G. The definition of a summary is a text which includes one or more words and these words represent important information in the raw text and shorter than the raw text obviously [20]. Table 1 presents the main different types of text summarization by Gambhir and Gupta [29]. The first type of summaries, amount of input document, is generated by single or multi-document. The second one has two kinds which is extraction method and abstract method. Extraction summarization is that extracts keywords and paragraphs to generate summaries. Abstraction summarization is that generates summaries by creating new texts. The third one is divided into two

kinds, generic and topic-oriented summarization. Generic summarization reflects the opinion of authors, while topic-oriented summarization is related to the topics which readers are interested in.

3. 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. The scope can be extended to abstractive summarization by including NLP features and implementing more scoring techniques. Use of semantic ranking can also be done to obtain meaningful summaries.
For the research in the future, we propose the following points:
 Extend the domains of essay and choose international journals that have higher rankings.

 Set more parameters and algorithms, such as attention mechanism. Limitation of the Study is decreasing the fluency of candidate titles and then evaluate with correct titles

4. REFERENCES

[1] Virat V. Giri, Dr.M.M. Math and Dr. U. P. Kulkarni , A Survey of Automatic Text Summarization System for Different Regional Languages in India Bonfring International Journal of Software Engineering and Soft Computing, Vol. 6, Special Issue, October 2016

[2] Sheetal Shimpikar and Sharvari Govilkar, A Survey of Text Summarization Techniques for Different Regional Languages in India, International Journal of Computer Applications, Vol. 165, No. 11, May 2017

[3] Sunitha C, Dr. A Jaya and Amal Ganesh, A Survey of Abstractive Summarization Techniques in Indian Languages, 2016

[4] Hamzah Noori Fejer and Nazlia Omar, Automatic Multi-Document Arabic Text Summarization Using Clustering and Keyphrase Extraction ICIMU IEEE 2014 International Conference,978-1-4799-5423-0.

[5] Deepali K. Gaikwad, Deepali Sawane and C. Namrata Mahender, Rule Based Question Generation for Marathi Text Summarization using Rule Based StemmerIOSR Journal of Computer Engineering (IOSR-JCE), e- ISSN: 2278-0661. 2015

[6] Mudassar Majgaonkar and Tanveer Siddiqui, Discovering suffixes: A case study for Marathi Language (IJCSE) International Journal on Computer Science and Engineering Vol. 02, No. 08, 2010, 2716-2720

[7] Aishwarya Sahani, Kaustubh Sarang, Sushmita Umredkar, and Mihir Patil, Automatic Text Categorization of Marathi Language Documents (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 7 (5) , 2016, 2297-2301

[8] Ms. Jayshri Arjun Patil, Ms. Poonam Bhagwandas Godhwani, Review of Name Entity Recognition in Marathi Language IJSART - Volume 2 Issue 6 , June 2016

[9] A report on Text Summarization for Compressed Inverted Indexes and snippets by Mahesh Dangale CS 297 Report July 2013.

[10] Feifan Liu, Yang Liu, Exploring Correlation between ROUGE and Human Evaluation on Meeting SummariesIEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING

[11] Pooja Pandey, Dhiraj Ahim, Sharvari Govilkar, Rule based Stemmer using Marathi WordNet for Marathi LanguageInternational Journal of Advanced Research in Computer and Communication Engineering, Volume 5, Issue 10, 2016

[12] Rafael Ferreira, Luciano de Souza Cabral, Assessing sentence scoring techniques for extractive text summarizationExpert Systems with Applications, Elsevier 2013

[13] TextRank: Bringing order into texts, Mihalcea, Rada., and Tarau, Paul. (2004), In Conference on empirical methods in natural language processing, Barcelona, Spain.

[14] Rouge: A package for automatic evaluation of summaries, Lin, C. Y. In Text summarization branches out, Proceedings of the ACL-04 workshop (Vol. 8).

[15] ”Variations of the Similarity Function of TextRank for Automated Summarization”, Federico Barrios, Federico Lopez, Luis Argerich, Rosita Wachenchauzer, arXiv, 2017.

[16] ”A Four Dimension Graph Model for Autom

[17] A. Sahba, J. Prevost Hypercube Based Clusters in Cloud Computing presented at 11th International Symposium on Intelligent Automation and Control, World Automation Congress 2016, Puerto Rico, July 2016

[18] A. Sahba, R. Sahba, and W.-M. Lin, ”Improving IPC in Simultaneous Multi-Threading (SMT) Processors by Capping IQ Utilization According to Dispatched Memory Instructions,” presented at the 2014 World Automation Congress, Waikoloa Village, HI, 2014.

[19] Erol, B. A., Vaishnav, S., Labrado, J. D., Benavidez, P., & Jamshidi, M. (2016, July). Cloud-based Control and vSLAM through cooperative Mapping and Localization. In World Automation Congress (WAC), 2016 (pp. 1-6). IEEE.

[20] Erol B.A., Majumdar A., Lwowski J., Benavidez P., Rad P., Jamshidi M. (2018) Improved Deep Neural Network Object Tracking System for Applications in Home Robotics. In: Pedrycz W., Chen SM. (eds) Computational Intelligence for Pattern Recognition. Studies in Computational Intelligence, vol 777. Springer, Cham.

[21] Amullen EM, Shetty S, Keel LH. Secured formation control for multi-agent systems under DoS attacks. In Technologies for Homeland Security (HST), 2016 IEEE Symposium on 2016 May 10 (pp. 1-6). IEEE.

[22] Amullen EM, Shetty S, Keel LH. Model-based resilient control for a multi-agent system against Denial of Service attacks. In World Automation Congress (WAC), 2016 2016 Jul 31 (pp. 1-6). IEEE.

[23] Farshid Sahba et al., ”Wireless Sensors and RFID in Garden Automation”, International Journal of Computer and Electronics Research, vol. 3, no. 4, August 2014.

[24] Farshid Sahba, Zahra Nourani, ”Smart tractors in pistachio orchards equipped with RFID”, presented at the 2016 World Automation Congress, 2016.

[25] H. Bouzary, F. Frank Chen, ”Service optimal selection and composition in cloud manufacturing: a comprehensive survey,” The International Journal of Advanced Manufacturing Technology, 2018.

[26] H. F. Azgomi, J. Poshtan, ”Induction motor stator fault detection via fuzzy logic”, Electrical Engineering (ICEE), 2013 21st Iranian Conference on , vol., no., pp.1,5, 14-16 May 2013.

[27] H. F. Azgomi, J. Poshtan, M. Poshtan, ”Experimental validation on stator fault detection via fuzzy logic”, 3rd international conf on EPECS, Istanbul, 2013.

[28] Dabbaghjamanesh, M., A. Kavousi-Fard, & S. Mehraeen. ”Effective Scheduling of Reconfigurable Microgrids with Dynamic Thermal Line Rating.” IEEE Transactions on Industrial Electronics (2018).

[29] Rakhshan, M., N. Vafamand, M. Shasadeghi, M. Dabbaghjamanesh, & A. Moeini. ”Design of networked polynomial control systems with random delays: sum of squares approach.” International Journal of Automation and Control 10, no. 1 (2016): 73-86.

[30] P. Shahmaleki, M. Mahzoon, and V. Shahmaleki, Designing Fuzzy Controller and Real Time Experimental Studies on a Nonholonomic Robot. IFAC Proceedings Volumes, 42(15), pp.312-319, 2009.

[31] P. Shahmaleki and M. Mahzoon, Designing a hierarchical fuzzy controller for backing-up a four wheel autonomous robot, 2008 American Control Conference, Seattle, WA, 2008, pp. 4893-4897.