A Survey of Text Summarization Techniques for Indian Regional Languages

Sheetal Shimpikar
PG Student
Department of Computer Engineering-PCE,
Mumbai University, India

Sharvari Govilkar
H.O.D
Department of Computer Engineering-PCE,
Mumbai University, India

ABSTRACT

Summarization is the process of reducing a text document with a computer program in order to create a summary that retains the most important points of the original document. Technologies that can make a coherent summary take into account variables such as length, writing style and syntax. The main idea of summarization is to find a representative subset of the data, which contains the information of the entire set. Text summarization is commonly used to handle summaries of email threads, action items from a meeting and simplifying text by compressing sentences used to manage knowledge and also to help Internet search engines. This paper gives comparative study of various text summarization techniques used for Indian regional languages and also discusses in detail two main Types of Text summarization techniques these are extractive and abstractive.

Keywords

NLP, text summarization, text summarization techniques, extractive, abstractive, features, Rich Semantic graph, TF-IDF, NLG.

1. INTRODUCTION

Text summarization is reducing a text with a computer program in order to create a summary that retains the most important points of original text. The main idea of summarization is to find a representative subset of the data, which contains the information of the entire set. Text summarizations choose the most significant part of text and create coherent summaries that state the main purpose of the given document. Text summarization can be categorised into given approaches. Single document summarizer means the summary is extracted from a single document, multiple document summarizer means the summary is extracted from a multiple document, generic document summarizer generates summaries containing main topics of document, query based document summarizer generate summaries containing sentences that are related to given queries . Abstractive and extractive summarization methods are used.

2. LITERATURE SURVEY

In this section we cite the relevant past literature of research work done in the field of text summarization techniques for Indian languages.

Sunitha.C, Dr.A.Jaya and Amal Ganesh worked on Abstractive summarization techniques in Indian languages. Authors have explained Abstractive summarization technique, classified in two approaches such as structure based approach and semantic based approach. There different methods are used in these approaches [1].

Jagadish S Kallimani suggests a solution for abstractive summarization by making use of extractive methodology in Kannada language. Performance accuracy for Literature 70%, for Entertainment 80%, for Sports 76%. The main idea is to generate abstractive summary by gathering key concepts from source document using extractive summary technique [2].

Manjula Subramanian discusses about semantic graph reducing technique to generate abstractive summary with input text in Hindi [3].

Rajina Kabeer used semantic Graph based method which concentrate on summarizing documents in Malayalam [4]. Atif Khan and Naomie Salim have worked on a paper of review on abstractive summarization methods [5].

Barzilay and Mckeown used Tree based techniques for text representation on dependency based representation: DSYNT tree. Content selection is theme intersection algorithm summary generation uses FUF/SURGE language generator [6].

Harabagiu and Lacatusu worked on template based method [7].

Lee and Jian worked on ontology based method, text representation was on Fuzzy ontology [8].

Green backer used multimodal semantic model [9].

Genest and Lapalme used INIT based method. Text representation on abstract representation information item [10].

Aadia Abbas Mohammed Elsied worked on automatic abstraction summarization a systematic literature review on abstractive summarization [11].

Rosna P Harun worked on text summarization methods in Dravidian language contains languages like Malayalam, Tamil, Telugu, Kannada, Kodagu, Badaga, Byari, and Tulu. [12].

Renjith SR have worked on automatic text summarization for Malayalam using sentence extraction [13].

Krish Perumal proposed a language independent sentence extraction based text summarization technique for English and Tamil [14].

Prajitha U proposed an algorithm namely LALITHA: A light weight Malayalam stemmer using suffix stripping method [15].

Pragisha K proposed a stemming algorithm namely STHREE: stemmer for Malayalam using three pass algorithm [16].

Dhanya P. M have done comparative study of text summarization in Tamil, Kannada, Odia, Bengali, Punjabi and Guajarati are taken for purpose of comparison [17]. Vishal Gupta worked on automatic Punjabi text extraction summarization system [18].

R.C.Balabantaray have worked on Odia text summarization using stemmer. A novel statistical approach has been applied to summarize the given Odia text [19].

Sankar K have worked on text extraction on Agglutinative language based on a graph theoretic model [20].

Kamal Sarkar worked on Bengali text summarization by sentence extraction [21].

Baxendale presented a straight forward method of sentence extraction for single document text summarization method for Bengali [22].

3. TEXT SUMMARIZATION PROCESS

The three important aspects which should be considered during text summarization are summary process may be result of single document or multiple documents, important information should be preserved, length of summary should be short. Text summarization can be decomposed into three phases, analysis phase, transformation phase and synthesis phase. In analysis phase, analyses input text and selects a few salient features. In transformation phase, transforms the result of analysis into summary representation. In synthesis phase takes summary representation and produces an appropriate summary corresponding to user needs. First is documents collection like .html, .pdf, .doc, web content. Second is pre-processing, which includes tokenization were **a** document is treated as a string, and then partitioned into a list of tokens. Stop word removal, stop words such as „a‟, „the‟, „and‟, and „at‟ etc. are removed. Stemming word, removing prefix and suffix of words reducing the words to their stem. Indexing, documents representation is one of the pre-processing technique that is to reduce the complexity of the documents and make them easier to handle, the document have to be transformed from the full text version to a document vector. Feature Selection to construct vector space, which improves the scalability, efficiency and accuracy of a text summarizer. Summarization Algorithm, the documents can be summarized by two ways, abstractive and extractive methods. Methods for summarization includes the machine learning approaches such as Bayesian classifier, Tree based, Ontology based, Rule based, K-Nearest Neighbour (K-NN), Support Vector Machine(SVMs), Neural Networks and more. Performance Evaluation, **t**his is last stage of text summarization, in which the evaluations of text summarizers is typically conducted experimentally, rather than analytically, like Precision and recall, fallout, error, accuracy etc.

Below is the diagram for text summarization process.

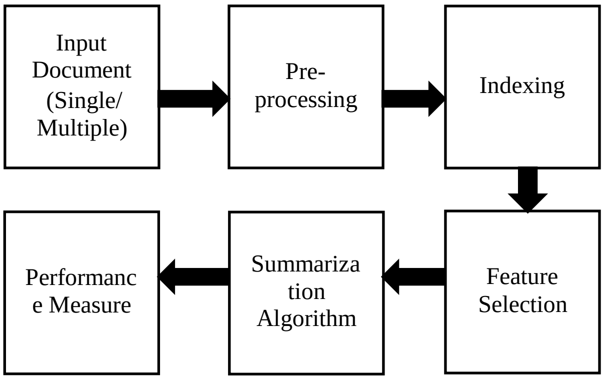

Figure 1. Text Summarization Process

4. TEXT SUMMARIZATION TECHNIQUES

Let us refer some of the techniques used in text summarization.

4.1 Abstractive Text Summarization Technique

Abstractive summarization techniques are classified into two categories. Structured based approaches and Semantic based approaches. Different methods are shown.

4.1.1 Structured based Approach

Structured based approach encodes most important information from the document(s), through cognitive schemas such as templates, extraction rules and other structures such as tree, ontology, lead and body phrase structure.

a. Tree based Method

This technique uses a dependency tree to represent the text/contents of a document. The technique uses either a language generator or an algorithm for generation of summary. In this approach, first the similar sentences are pre-processed using a shallow parser and then sentences are mapped to predicate-argument structure. Next, the content planner uses theme intersection algorithm to determine common phrases by comparing the predicate argument structures. Those phrases that convey common information are selected and ordered and some information are also added with it (temporal references, entity descriptions).Finally sentence generation phase uses FUF/SURGE language generator to combine and arrange the selected phrases into new summary sentences. The major strength of this approach is that the use of language generator significantly improved the quality of resultant summaries i.e reducing repetitions and increasing fluency. The problem with this approach is that context of sentence was not included while capturing the intersected phrase.

b. Template based Method

This technique uses a template to represent a whole document. Linguistic patterns or extraction rules are matched to identify text snippets that will be mapped into template slots. These text snippets are indicators of the summary content. This method is used for Kannada language.

c. Ontology based Method

Used to improve the process of summarization. In this method, domain ontology for news event is defined by the domain experts. Next phase is document processing phase. Meaningful terms from corpus are produces in this phase. The meaningful terms are classified by the classifier on basis of events of news. Membership degree associated with various events of domain ontology. Membership degree is generated by fuzzy inference. Limitations of this approach are it is time consuming because domain ontology has to be defined by domain experts. Advantage of this approach is it handles uncertain data.

d. Rule based Method

The rule based method comprises of three steps. Firstly, the documents to be classified are represented in terms of their categories. The categories can be from various domains. Hence the first task is to sort these. The next thing is to form questions based on these categories. E.g. amongst the various categories like attacks, disasters, health etc., taking the example of an attack category several questions can be figured out like: - What happened? , when did it happen? Who got affected? What were the consequences? Etc. -Depending upon these questions, rules are generated. Here several verbs and nouns having similar meanings are determined and their positions are correctly identified. -The context selection module selects the best candidate amongst these. -Generation patterns are then used for the generation of summary sentences.

4.1.2. Semantic based Approach

In Semantic based method, semantic representation of document(s) is used to feed into natural language generation (NLG) system. This method focus on identifying noun phrases and verb phrases by processing linguistic data. Different methods using this approach are discussed here.

a. Multimodal Semantic Model

In this method, a semantic model, which captures concepts and relationship among concepts, is built to represent the contents (text and images) of multimodal documents. The important concepts are rated based on some measure and finally the selected concepts are expressed as sentences to form summary.

b. Information Item based Method

In this method**,** the contents of summary are generated from abstract representation of source documents, rather than from sentences of source documents. The abstract representation is Information Item, which is the smallest element of coherent information in a text. From this method, a short, coherent, information rich and less redundant summary can be formed. In spite of so many advantages, this method has also many limitations. While making grammatical and meaningful sentences, many important information items get rejected. Due to which, linguistic quality of resultant summary gets reduced.

c. Semantic graph based Method

This method aims to summarize a document by creating a semantic graph called Rich Semantic Graph (RSG) for the original document, reducing the generated semantic graph, and then generating the final abstractive summary from the reduced semantic graph. The main objective of this method is generating a summery by creating a semantic graph called rich semantic graph (RSG). The semantic graph approach consists of three phases. The first phase represents input document using rich semantic graph, the verbs and nouns of the input document are represented as graph nodes and the edges correspond to semantic and topological relations between them. The second phase reduces the original graph to a more reduced graph using heuristic rules. The third phase generates an abstractive summery. The advantage of this method is that it produces less redundant and grammatically correct sentences. The disadvantage of this method is that it is limited to a single document and not multiple documents. Hindi and Malayalam and Tamil languages are using this method.

4.2 Extractive Text Summarization Technique

In extractive-based methods, there are different approaches to implement extractive summaries.

4.2.1. Term frequency- Inverse Document Frequency (TF-IDF)

It is a word distribution method in which two measures namely term frequency (tf) and document frequency (df) are calculated for each non-stop-word (w) in the document. Term frequency (tf), indicate number of times a word appears in the text which measures salience of word within that document. Document frequency (df) indicate number of documents in which the word appears. The frequent occurrence of a word in a document is treated as informative word, which is calculated by document frequency measure. Thematic words are obtained by comparing the ratio between two frequencies, referred as (if*idf) measure. Once (tf-idf) score has been computed for each word the next step is to calculate number of such thematic words per sentence. With this value sentences in the input text are ranked and highest scored sentences are picked to be part of summary. Redundancy of information is extremely high in this method. Punjabi, Bengali, Kannada, Odia languages are using this method.

4.2.2. Cluster based Method

Measures relevance or similarity between each sentence in a document with that of sentences selected for summary. Summaries address onto different “themes” appearing in the documents, which is incorporated through clustering. Clustering based methods become essential to generate a meaningful summary. Several clustering techniques are also available for text categorization like K-means, Suffix tree Clustering, Label Induction Grouping (LINGO) Algorithm, Semantic Online Hierarchical Clustering (SHOC). Clustering of documents is mainly used to minimize the amount of text by categorizing or grouping similar data items. The following is the brief introduction of some clustering approaches.

a. K- means Algorithm

This algorithm is an iterative algorithm where the number of input clusters is needed to be mentioned. In this algorithm dataset is split into K clusters and the data points are arbitrarily assigned to the resulting clusters that have roughly the similar number of data points. For each data point the difference from the data point to each cluster is evaluated. If the data point is closer to its own clusters than keep it as it is. Suppose if the data point is not closer to its cluster, copy it into the nearest cluster. The advantage is if the clusters are global than it produces tighter clusters than hierarchical clustering.

b. Maximum marginal relevance multi document (MMR-MD)

Summarization is a cluster based extractive summarization method. This method aims at generating summary of high relevance, to the given query or document topic. Passage clustering forms a main component in this system, to extract most relevant sentences of a documents by keeping the summary non-redundant. Similarity measure is used to initially cluster given document or documents. This ensures that most representative sentences of the document are chosen, while ensuring minimum redundancy in the summary.

c. Graph Theoretic

Representation is an extractive summarization model, which provides a method to identify themes in the document. Pre-processing steps, namely, stop word removal and stemming are done before, to obtain graphical view of the documents. Graphic representation yields partitions indicating distinct topics covered in the documents. Identification of nodes with high cardinality forms higher preferred sentences to be included in the summary. Tamil language summarization system for scoring of sentences in summary using graph theoretic scoring technique. This system uses statistics of frequency of words and a term positional and weight-age calculation by string pattern for scoring of sentences.

5. CONCLUSIONS

A brief summaries of automatic text summarization techniques for various Indian regional languages has been described in the document. We can notice that good work has been done for various languages like Bengali, Malayalam, Hindi, Odia, Gujarati, Tamil, Telugu, Gujarati and Punjabi etc. 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 Marathi sentences. Also, we will try different machine learning techniques for comparison and try to achieve better accurate results. Several text summarization algorithms were proposed for the automatic summarization of documents. Among these algorithms Rich semantic graph based, Rule based, Tree based, Ontology based, TFIDF and cluster based techniques are shown most appropriate in the existing literature. After performing a review on different types of approaches and comparing existing methods based on various parameters it can be concluded that TFISF, Graph based are recognized as most effective text summarization method for Indian Regional Languages.

6. ACKNOWLEDGEMENTS

I am using this opportunity to express my gratitude to thank all the people who contributed in some way to the work described in this paper. My deepest thanks to my project guide for giving timely inputs and giving me intellectual freedom of work. I express my thanks to the head of computer department and to the principal of Pillai Institute of Information Technology, New Panvel for extending his support.

7. REFERENCES

[1] Sunitha C, Dr. A Jaya and Amal Ganesh “Abstractive Summarization Techniques in Indian Languages”, Peer-review under responsibility of the Organizing Committee of ICRTCSE 2016 doi: 10.1016/j.procs.2016.05.121, International Conference of recent trends in computer science.

[2] Jagadish S Kallimani, “Summarizing News Paper Articles: Experiments with Ontology- Based, Customized, Extractive Text Summary and Word Scoring”, Research Scholar, Department of Computer Science and Engineering, Jawaharlal Nehru Technological University, Kakinada, Andra Pradesh, India, 2012.

[3] Manjula Subramanian, “Semantic graph reducing technique to generate abstractive summary with input text in Hindi.

[4] Rajina Kabeer, “Semantic Graph based method on summarising documents in Malayalam”, IEEE International conference on data science and engineering, 2014.

[5] Atif Khan and Naomic Salim, “Abstractive Summarization Methods”, Journal of Theoretical and applied information technology, vol.55, ISSN- 1992-8645, E-ISSN-1817, 3195, September-2013.

[6] Barzilay and Mckeown, “Sentence fusion for multi document news summarization”, Computer Linguistics, vol-31, pp. 297-338, 2005

[7] Harabagin and Lacatusu, “Generating single and multi-document summaries with gistexter”, in document understanding conference, 2002.

[8] Lee and Jian, “Ontology based method, text representation on Fuzzy ontology and content selection is classifier”, Systems, Man and Cybernetics, Part B: Cybernetics, IEEE Transaction on, vol.35, pp. 859-880, 2005

[9] Green backer, “Multimodal semantic model”, ACL HLT 2011, pp.75, 2011.

[10] Genest and Lapalme, “Framework for abstractive summarization using text to text generation”, in Proceeding of workshops on Monolingual Text to Text generation., 2011, pp.64-73, information item.

[11] Hadia Abhas Mohammed Elsied, Naomic Salim, “Automatic Abstraction Summarization a systematic Literature review”, Journal of Theoretical and Applied Information Technology 31st August 2013. Vol. 54 No.3, ISSN: 1992-8645, E-ISSN: 1817-3195.

[12] Rosna P Harun, “Text Summarization methods in Dravidian Language”, International Journal of Innovations in Engineering and Technology (IJIET), Volume 7 Issue 1 June 2016.

[13] Renjith SR and Sony P, “Automatic text summarization for Malayalam using sentence extraction”. Proceedings of 27th IRF International Conference, 14th June 2015, Chennai, India, ISBN: 978-93-85465-35-2

[14] Krish Perumal, Bidyut baran Chaudhuri Language Independent Sentence Extraction based Text Summarization, In Proceedings of ICON 2011, and 9th International Conference on Natural Language Processing.

[15] Prajitha, “An algorithm namely LALITHA-A, A light weight Malayalam stemmer”, Control Communication and Computing (ICCC), 2013 International Conference on, **DOI:** 10.1109/ICCC.2013.6731658, 06 February 2014.

[16] Pragisha K. and P.C. Raghu Raj. STHREE: Stemmer for Malayalam using THREE pass algorithm. Presented in the IEEE ICCC 2013, College of Engineering, and Trivandrum.

[17] Dhanya P M and Jathavedam M, "Comparative study of text summarization in Indian Languages", International Journal of Computer Applications (0975 – 8887) Volume 75– No.6, August 2013.

[18] Vishal Gupta and Gurpreet Singh Lehal, "Automated Punjabi text extraction summarization system", Manuscript received January 12, 2010; revised March 22, 2010; Accepted April 29, 2010, doi:10.4304/jetwi.2.3.258-268

[19] R.C. Balabantaray, B Sahoo and Mswan,"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 – www.ijais.org

[20] Sankar K, Vijay Sundar Ram K and Sobha Devi, "Text extraction on Agglutinative languages", Language in India www.languageinindia.com 1 1: 5 May 2011.

[21] Kamal Sarkar, "Bengali text summarization by sentence extraction", Computer Science & Engineering Department, Jabalpur University, Das, A. & Bandyopadhyay, S. 2010. Topic- Based Bengali Opinion Summarization. COLING (Posters) 2010: 232-240.

[22] Baxendale, "Straight forward method of sentence extraction using document title", Proceedings of the NAACL HLT Workshop on Integer Linear Programming for Natural Language Processing, pages 1–9, Boulder, Colorado, June 2009. C 2009 Association for Computational Linguistics.

[23] N. R. Kasture1, Neha Yargal 2, Neha Nityanand Singh3, Neha Kulkarni4 and Vijay Mathur5, "A Survey on Methods of Abstractive Text Summarization", International journal for research in emerging science and technology, volume-1, issue-6, November-2014.

[24] tf–idf https://en.wikipedia.org/wiki/Tf%E2%80%93idf