Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 87 (2016) 25 – 31

Procedia Computer Science

Fourth International Conference on Recent Trends in Computer Science & Engineering

Chennai, Tamil Nadu, India

A Study on Abstractive Summarization Techniques in Indian Languages

Sunitha.C[a]*, Dr.A.Jaya[b], Amal Ganesh[a]

[a]*Department of CSE, Vidya Academy of Science and Technology, Thrissur, Kerala, 680501, India*

[b]*Department of MCA, B.S Abdur Rahman University, Vandalur, Chennai, 600048, India*

Abstract

Natural Language Processing is a vast area which has great importance when people started to interpret human language from one form to another. Summarization is one of the research works in NLP which concentrates on providing meaningful summary using various NLP tools and techniques. Since huge amount of information is used across the digital world, it is highly essential to have automatic summarization techniques. Extractive and Abstractive summarization are the two summarization techniques available. A lot of research works are being carried out in this area especially in extractive summarization. Even though more works are carried out using extractive method, meaningful summary can be attained using abstractive summary techniques which make it more complex. In Indian languages, very few works are carried out in the field of abstractive summarization and there is high need for having research works being carried out in this area. Here, we are concentrating on the various techniques available for abstractive summarization and also try to explain the limited works currently available in abstractive summary field of Indian languages.

Keywords: Abstractive Summary, Semantic Graph, Ontology

1. Introduction

* Corresponding author. Tel.: 0091-9446076935
E-mail address: sunitha@vidyaacademy.ac.in

Peer-review under responsibility of the Organizing Committee of ICRTCSE 2016
doi:10.1016/j.procs.2016.05.121

Natural Language Processing [NLP] gained its attention when humans started to interpret one form of language to machine language. A lot of research works are being carried out in the field of NLP. In this modern era, data retrieval across websites and other informative medias are used everywhere irrespective of the languages we speak which made inevitable for having NLP applications like summarization [1][2][3]. Summarization is the process of abstracting the information pertaining to a particular area/domain/topic from various reliable sources. Summarization technique is very handy in various applications such as online new articles summary, product review summary, one line email summary, automated research abstract, abstracted information summary for government officials, business organization etc. with minimum human intervention [4].

Summarization techniques are classified into two categories namely, extractive summarization technique and abstractive summarization technique [5][6]. Extractive summary technique involves generating summary based on the key features in the given text. It makes use of statistical methods such as term frequency, location, cue method, title/headline word, sentence length, similarity, proper noun, proximity etc. to find the sentences required for generating summary. The summary generated using extractive method will have the same sentences abstracted from source text which makes its implementation little bit easier. Majority of research works are carried out in extractive methods [7] but the summary generated need not be complete as well as meaningful. Whereas in abstractive summary technique, meaningful summary is generated in the same way as it is generated by humans. The abstracted summary may contain words/sentences that are not part of the source text. In abstractive method, the complexity arises due to the main facts such as how to select the important concepts from the text without losing its meaning; how to represent it in a condensed manner; and how to generate it in a reproducible manner. Some of the existing approaches are sentence compression, sentence fusion and sentence splitting. But these are the works carrying over the existing extractive approaches. But it would be better to use the linguistic approaches to get the meaningful summaries for Indian languages.

In Indian languages, the research works carried out in abstractive methods are nominal which motivated us to write this paper. In this paper, we will be concentrating on the abstractive summary techniques used in Indian languages. This paper is divided in to three parts namely; different methods in abstractive summarization, related woks carried out in Indian languages, comparative analysis of the methods and finally conclusion.

2. Abstractive Summarization Methods

The abstractive summarization technique is broadly classified in two approaches such as Structure based approach and Semantic based approach [8].

2.1. Structure based approach

In structure based approach, important sentences from source text gets populated in a predefined structure to obtain the required abstract summary without losing its meaning. The predefined structures used in this approach are templates, tree-based structure, ontology based structure, lead and body phrase structure and rule based structure. In template based method, from the snippets extracted using keywords, the required information are populated into a template to form the final summary [9][10]. In tree based method similar sentences are extracted from source text with the help of a parser and populated into a tree structure which follows predicate-argument structure[11[12]. In ontology based method the source text is preprocessed to extract the required keywords which are mapped as concepts and relations with the help of predefined ontology which will be converted to meaningful summary [13]. The lead and body phrase method focuses on revamping the lead sentence by either substituting or inserting the information rich similar phrases from the body which are called triggers.If the body phrase has higher syntactic similarity with the lead phrase, then that lead phrase gets substituted by the body phrase provided the information is richer than the lead phrase. Insertion happens if there is no similarity between body and the lead phrases [14]. In rule based method rules and categories are used to represent the document summaries. Rules are fed into this module to get the required meaningful candidates from which the best candidate is selected which is passed to summary generation module. Finally, the summary is generated using generation pattern [15].

2.2. Semantic based approach

Semantic based approach involves three stages namely; inputting the document, semantic representation of the document and feeding semantic representation to Natural Language Generation phase (NLG) to obtain the desirable output. The different methods used in this approach are Multimodal semantic model, Information item based method and Semantic Graph based method. In multimodal semantic model the semantic model is constructed by making use of concepts and finding the relation between these concepts with the help of ontology. The second stage involves identifying the important concepts using information density metrics. In the final stage, the required summary is generated from these important concepts [16]. The information item based method deals with abstract representation of the document to form information items. These information items are extracted after syntactic analysis of the text. From these items, sentences are generated by obeying the subject-verb-object structure using a sentence generator. The sentences generated are then ranked based on the average Document Frequency (DF) score. From this list, the highly ranked sentences are taken to form the summary [17]. The semantic graph based method consists of three phases. In the first phase, the entire document is represented by a Rich Semantic Graph (RSG). In the second phase, heuristic rules are applied to reduce the complexity of semantic graph. Then in the final phase, the abstract summary is generated from this reduced graph [18][19].

3. Related works in Indian Languages

The abstractive summarization research works in Indian languages are in premature state when compared to other languages like English. French, Arabic, Spanish, Chinese, German etc. This is mainly due to the diversity in Indian languages and the lack of resources such as raw data, various NLP tools etc[20]. This section explains the very few abstractive summarization works in Indian languages like Telugu, Hindi, Bengali, Kannada and Malayalam.

Jagadish S Kallimani, et al suggests a solution for abstractive summarization by making use of extractive methodology in Kannada language [21]. In this method, the abstract data is extracted from the source document. This information dense data is then post processed to gather the key or most important concepts from original text. The main idea is to generate abstractive summary by gathering key concepts from source document using extractive summary technique.

This technique involves three phases such as pre-processing, summarizing and post processing as shown in Figure-1. In pre-processing, various processes like Data Chunking, Stemming, Named Entity Recognition, Word Sense Disambiguation, Stop word removal, Parts of Speech Tagging etc. are applied on the input text. The next phase is summarizer module which makes use of two main components namely word cues and word snippets for extracting the key features available in input document. In the final phase, post processing is done by modifying the extractive summary to abstractive summary. This is done by undergoing refinement and rephrasing of the summary currently obtained. Finally, this well-structured abstractive summary is inputted to Text to Speech (TTS) engine where it finds useful in different real-time applications like news summarization and remote call handling.

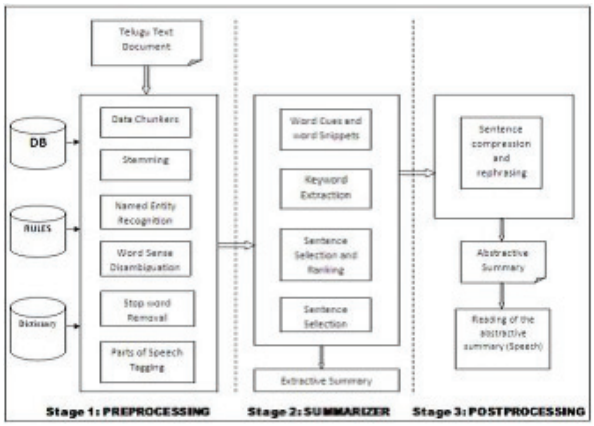

Fig- 1 : Abstractive summarization process

In his another paper, Jagadish S Kallimani, et al proposed a topic based guided summarization technique where rules are applied to generate an abstractive summary from a document [22, [23]. This technique was initially implemented in Kannada language and later got successfully reproduced in Hindi, Bengali and Telugu languages.

This method has mainly 4 modules as shown in Figure-2 namely; pre-processing module, categorization module, attribute extraction module, and summary generation module. In pre-processing module, the lemmatization, stemming, parts of speech tagging etc are applied to the input document. In categorization module, the pre-processed document is given as input to a classifier which will identify the category of the document. The next stage involves attribute extraction where content relevant attributes are extracted from category identified document. The attributes are then mapped to a template to get the required abstractive summary. The monotony in the structure of the summary generated sentences can be avoided by making use of tools like WordNet and SimpleNLG[24].

Fig 2: The abstractive summarization system for Kannada

Manjula Subramanian et al discusses about semantic graph reducing technique to generate abstractive summary with input text in Hindi [25].This approach is divided into three phases as shown in Figure-3; developing RSG from source document, reducing RSG to form the abstractive RSG and the final step includes the generation of abstractive summary from abstracted RSG.

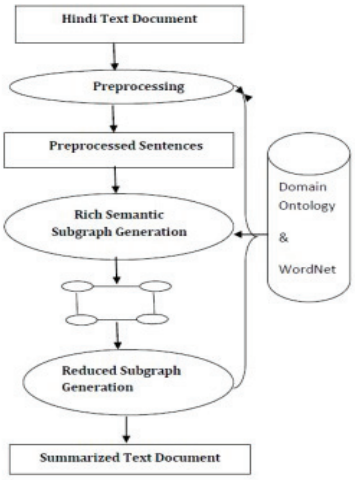

Fig 3: Semantic graph based approach for Hindi

The first phase starts with deep understanding of the grammatical (syntactic) structure of the input text. In the next stage, the model obtained will map with domain-based ontology to redefine the graph in terms of concepts and relations. These concepts and relations are mapped to form the final Rich Semantic Graph (RSG). This Rich Semantic Graph needs to be reduced to remove redundancy in concepts and relations represented. This can be achieved by applying heuristics rules using WordNet relations to reduced RSG. The reduced RSG needs to be represented in a summarized version by mapping it with domain-based ontology forming multiple sentences which are ranked based on relevance to the topic. Based on the ranking, the sentences are selected to form the final abstractive summary.

Another work on Semantic Graph based method is proposed by Rajina Kabeer et al which concentrate on summarizing documents in Malayalam [26]. In this method, the input document undergoes series of linguistic processing to get the triples from each sentence of the source document as shown in Figure - 4.With the help of these semantic triples, the semantic graph is generated and the graph needs to be reduced in order to remove redundancy and to attain concise abstract summary.

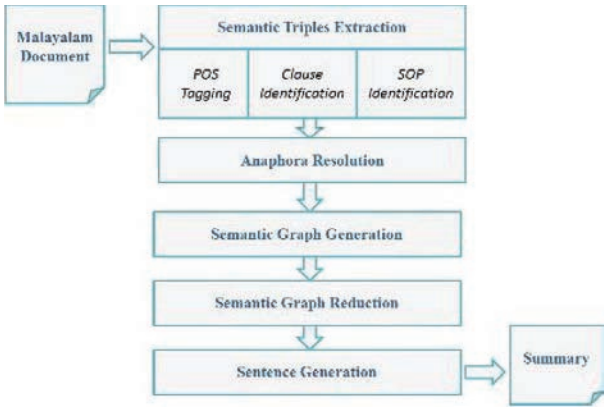

Fig 4: Semantic graph based approach for Malayalam

4. Comparison of abstractive methods in Indian languages

Based on these related works, an analytical comparison is drawn between these abstractive summarization methods in Indian languages. The Table-1 shows the different abstractive methods used in Indian languages and their performance measures. The analysis is done based on the following parameters such as the type of abstractive summarization method used, the type of data set used for performance measure and the primary performance attributes such as Precision, Recall and F-measure [27].

Table-1: Comparison of abstractive methods in Indian languages

	Abstractive Method used	Data set	Precision	Recall	F-measure
Hindi	Semantic graph method	NA	NA	NA	NA
Malayalam	Semantic graph method	Rouge	0.466	0.667	0.400
Kannada	Sentence rephrasing on extracted summary	NA	NA	NA	NA
Kannada	Template based method	30 documents across six different categories	0.8624	0.7893	0.815

5. Conclusion

The enormous amount of digital data available which is increasing day-by-day make the summarization inevitable in most of the application areas like news summarization, summary abstract of emails, scholarly journals, papers etc. Extractive and abstractive methods account for the two summarization techniques available in NLP. Even though extractive summarization works are more common in this area, a more precise, complete and meaningful summary can only be obtained using abstractive summary technique. Also, the abstractive works carried out in Indian languages are very limited compared to other countries. Here, in our paper we concentrate on different abstractive summarization techniques available and the various works currently available in Indian languages. This paper will give the readers a clear idea of the scarcity of the work done in the field of abstractive summarization in Indian languages. Also, it will give us the opportunity and confidence to move forward with abstractive summarization techniques using various methods like domain-based ontology, semantic graphic representation, WordNet etc.

References

1. Dipanjan Das, Andre F.T.Martins, “A survey on automatic text summarization”, Language Technologies Institute, Carnegie Mellon University, 2007
2. H. Saggion and T. Poibeau, "Automatic text summarization: Past, present and future," in Multi-source, Multilingual Information Extraction and Summarization, ed: Springer, 2013, pp. 3- 21.
3. M. Haque, et al., "Literature Review of Automatic Multiple Documents Text Summarization," International Journal of Innovation and Applied Studies, vol. 3, pp. 121-129, 2013.
4. Jagadish S KALLIMANI, Srinivasa K G and Eswara REDDY B, "Summarizing News Paper Articles:Experiments with Ontology-Based, Customized, Extractive Text Summary and Word Scoring" , in Cybernetics and Information Technolgies, Vol 12, No.2.
5. Neelima Bhattia and Arunima Jaiswal, “Trends in extractive and abstractive techniques in text summarization”, International Journal of Computer Applications(0975-8887), Volume 117-No.6, May 2015.Vishal Gupta and Gurpreet Singh Lehal , “A Survey of Text Summarization Extractive Techniques”, in the Journal of Emerging Technologies in Web Intelligence, Vol, 2, No. 3, August 2010.
6. Vishal Gupta, “A survey of text summarizers for Indian Languages and comparison of their performance” in journal of emerging technologies in web intelligence, vol 5, No. 4, November 2013.
7. Atif Khan and Naomie Salim, “A review on abstractive summarization techniques,” Journal of theoretical and applied information technology vol. 55, No. 3. ISSN: 1992-8645, E-ISSN: 1817-3195, September 2013.
8. Tatsuro Oya, Yashar Mehdad, Giuseppe Carenini, Raymond Ng , “A Template-based Abstractive Meeting Summarization: Leveraging Summary and Source Text Relationships”, Proceedings of the 8th International Natural Language Generation Conference, pages 45–53, Philadelphia, Pennsylvania, 19-21 June 2014. c 2014 Association for Computational Linguistics.
9. S. M. Harabagiu and F. Lacatusu, "Generating single and multi-document summaries with gistexter," in Document Understanding Conferences, 2002.
10. R. Barzilay and K. R. McKeown, "Sentence fusion for multidocument news summarization," Computational Linguistics, vol. 31, pp. 297-328, 2005.

11. R. Barzilay, et al., "Information fusion in the context of multi-document summarization," in Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics, 1999, pp. 550- 557.
12. C.-S. Lee, et al., "A fuzzy ontology and its application to news summarization," Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on, vol. 35, pp. 859-880, 2005.
13. H. Tanaka, et al., "Syntax-driven sentence revision for broadcast news summarization," in Proceedings of the 2009 Workshop on Language Generation and Summarisation, 2009, pp. 39-47.
14. P.-E. Genest and G. Lapalme, "Fully abstractive approach to guided summarization," in Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers- Volume 2, 2012,pp.354-358.
15. C. F. Greenbacker, "Towards a framework for abstractive summarization of multimodal documents," ACL HLT 2011, p. 75, 2011.
16. P.E. Genest and G. Lapalme, "Framework for abstractive summarization using text- to-text generation," in Proceedings of the Workshop on Monolingual Text-To-Text Generation, 2011, pp. 64-73.---information item.
17. I. F. Moawad and M. Aref, "Semantic graph reduction approach for abstractive Text Summarization," in Computer Engineering & Systems (ICCES), 2012 Seventh International Conference on, 2012, pp. 132-138.
18. 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.
19. Dhanya P M and Jathavedan M, “Comparative study of text summarization in Indian languages”, in International Journal of Computer Applications (0975 – 8887) Volume 75– No.6, August 2013
20. Jagadish S KALLIMANI, Srinivasa K G and Eswara REDDY B, “Information Extraction by an Abstractive Text Summarization for an Indian Regional Language”, IEEE 2011.
21. Jagadish S KALLIMANI, Srinivasa K G and Eswara REDDY B,”A comprehensive analysis of guided abstractive summarization”, in International Journal of Computer Science Issues, Volume 11, Issue 6, No 1, November 2014, ISSN (Print): 1694-0814 | ISSN (Online): 1694-0784.
22. Varsha R Embar, Surabhi R Deshpande, Vaishnavi A K, Vishakha Jain, Jagadish S Kallimani, “ sArAmsha - a Kannada Abstractive Summarizer”, 978-1-4673-6217-7/13/$31.00 c 2013 IEEE .
23. A. Gatt and E. Reiter, "SimpleNLG: A realisation engine for practical applications," in Proceedings of the 12th European Workshop on Natural Language Generation, 2009, pp.90-93.
24. Manjula Subramaniam and prof. Vipul Dalal, “Test Model for Rich Semantic Graph Representation for Hindi Text using Abstractive Method” in Volume: 02 Issue: 02 | May-2015, e-ISSN: 2395 -0056, p-ISSN: 2395-0072.
25. Rajina Kabeer and Sumam Mary Idicula, “Text Summarization for Malayalam documents – An experience”, in IEEE International Conference on Data Science & Engineering (ICDSE), 2014..
26. Josef Steinberger, Karel Je˘zek, “Evaluation measures for text summarization” in Computing and Informatics, Vol. 28, 2009, 1001–1026, V 2009-Mar-2.