Text Summarization Using Abstractive Methods

Pankaj Rawat
Department of Computer Applications, National Institute of Technology, Kurukshetra, Haryana, India.

Dr. Nikam Gitanjali Ganpatrao
Assistant Professor, Department of Computer Applications, National Institute of Technology, Kurukshetra, Haryana, India.

Deepak Gupta
Department of Computer Applications, National Institute of Technology, Kurukshetra, Haryana, India.

Abstract – Text summarization is the procedure of extracting vital and important information from the given source text and to produce that information to the user in the form of a summary. Sometimes it becomes very difficult for humans to manually summarize a large document of text. Automatic abstractive text summarization provides the required solution but it not an easy task as it requires a deeper analysis of given text document. Through this paper, we are presenting a survey on abstractive text summarization methods. Abstractive methods are broadly classified into two categories namely, structured based approach and semantic based approach. Advantages and disadvantages of each method are also highlighted. Lastly, it is concluded from the literature studies that most of the abstractive text summarization methods produce highly coherent, cohesive, information rich and less redundant summary.

Index Terms – Abstractive Summary, Extractive Summary, Semantic Graph, Abstraction Scheme, Sentence.

1. INTRODUCTION

There is no denying the fact that data on the Internet is growing at an exponential pace. Nowadays, people use the internet to search for the information through Information Retrieval (IR) tools such as Google, Yahoo, Bing etc. Information abstraction or summary of the retrieved result has become a necessity for the users. In the current era of information overload, text summarization has become an important and timely tool for a user to quickly understand the large volume of the information. The main goal of an automatic text summarization is compressing a document into a shorter version and yet preserving most of the important contents (if not all).

A summary helps the user for finding the key information in the document. For humans, generating a summary is a very straightforward process but it is very time consuming. Therefore, the need for an automated summary generation is becoming more and more apparent to get the general idea of long textual data.

So what exactly is the important information in a document? Finding out the important information is a truly challenging task. The need for an automatic text summarization is apparent in areas such as news articles summary, short message news on mobile, email summary and summary of chapters from text books. The first effort on automatic text summarization system was made in the late 1950. This automatic summarizer selects significant sentences from the document and concatenates them together. The approach in [1] uses term frequencies to measure sentence relevance and sentences with higher term frequencies were included in the summary.

Text summarization approaches are divided into two groups namely extractive and abstractive summarization. Extractive summarizations extract important sentences or phrases from the original documents and group them to produce a summary without changing the original text. An extractive text summarization system is proposed based on POS tagging by considering Hidden Markov Model using corpus to extract important phrases to build as a summary [2]. Abstractive summarization consists of understanding the source text by using linguistic method to interpret and examine the text. Abstractive methods need a deeper analysis of the text. These methods have the ability to generate new sentences, which improves the focus of a summary, reduce its redundancy and keeps a good compression rate [3].

We have presenting a survey on abstractive text summarization methods. Abstractive methods are broadly classified into two categories namely, structured based approach and semantic based approach. Advantages and disadvantages of each method are also highlighted. Lastly, it is concluded from the literature studies that most of the abstractive text summarization methods produce highly coherent, cohesive, information rich and less redundant summary.

2. TEXT SUMMARIZATION FEATURES

Text summarization identifies and extracts key sentences from the source text and concatenates them to form a concise summary. In order to identify key sentences for summary, a list features as discussed below, can be used to for selection of key sentences.

Term Frequency: Statistics provide salient terms based on term frequency, thus salient sentences are the ones that contain the words that occur frequently [4].The score of sentences increases for each frequent word. The most common measure widely used to calculate the word frequency is TF IDF.

Location: It relies on the intuition that important sentences are located at certain position in text or in paragraph, such as beginning or end of a paragraph [5].

Cue Method: Words that would have positive or negative effect on the respective sentence weight to indicate significance or key idea [5] such as cues: "in summary", "in conclusion", "the paper describes", "significantly".

Title/Headline word: It assumes that words in title and heading of a document that occur in sentences are positively relevant to summarization [4].

Sentence length: Short sentences express less information and therefore excluded from summary. Keeping in view the size of summary, very long sentences are also not appropriate for summary [5].

Similarity: This feature determines similarity between the sentence and the rest of the document sentences and similarity between the sentence and title of the document. Similarity can be calculated with linguistic knowledge or by character string overlap [4].

Proper noun: Sentences having proper nouns are considered important for document summary. Examples of proper nouns are: name of a person, place or organization [5].

Proximity: The distance between text units where entities occur is a determining factor for establishing relations between entities [5].

3. ABSTRACTIVE SUMMARIZATION APPROACH

Summarizations using abstractive techniques are broadly classified into two categories: Structured based approach and Semantic based approach [3].

1) Structured Based Approach

Structured based approach encodes most important information from the document through cognitive schemes such as templates, extraction rules and other structures such as tree, ontology, rule based structure [3]. Brief abstract of all the techniques under structure based approach is provided in Table 1.

2) Semantic Based Approach

In Semantic based approach, semantic representation of document is used to feed into natural language generation (NLG) system. This method focuses on identifying noun phrase and verb phrase by processing linguistic data [3]. Brief abstract of all the techniques under semantic based approach is provided in Table 2.

3.1 Brief discussion on structure based abstractive text summarization

3.1.1. Rule based method

In this method, the documents to be summarized are represented in terms of categories and a list of aspects. Content selection module selects the best candidate among the ones generated by information extraction rules to answer one or more aspects of a category. Finally, generation patterns are used for generation of summary sentences. The methodology in [6] generates short and well written abstractive summaries from clusters of news articles on same event. The methodology is based on an abstraction scheme. The abstraction scheme uses a rule based information extraction module, content selection heuristics and one or more patterns for sentence generation. Each abstraction scheme deals with one theme or subcategory. In order to generate extraction rules for abstraction scheme, several verbs and nouns having similar meaning are determined and syntactic position of roles is also identified. The information extraction (IE) module finds several candidate rules for each aspect of the category. Based on the output of the IE module, the content selection module selects the best candidate rule for each aspect and passed it to summary generation module. This module form summary of text using generation patterns designed for each abstraction scheme. The strong point of this method is that it has a potential for creating summaries with greater information density than current state of art. The main drawback of this methodology is that all the rules and patterns are manually written, which is tedious and time consuming.

3.1.2. Ontology Method

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 [7].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.

3.1.3. Tree Based Method

In this approach, the preprocessing is done of similar sentences using shallow parser [8]. After that we map those sentences to the predicate-argument structure. Different algorithms can be used for selecting the common phrase from the sentences such as Theme algorithm. The phrase conveying the same meaning is selected and also we add some information to it and will

arrange in a particular order. At the end, FUF/SURGE language generator can be used for making the new summary sentences by combining and arranging the selected common phrase. Use of language generator increases the fluency of the language and also reduces the grammatical mistakes. This feature is the main strength of this method. The main problem with this method is that the context of the sentences does not get included while selection of common phrase and it is important part of the sentences even if it is not part of the common phrase.

3.2 BRIEF DISCUSSION ON SEMANTIC BASED ABSTRACTIVE TEXT SUMMARIZATION

3.2.1. 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. In [12], a framework was proposed for generating an abstractive summary from a semantic model of a multimodal document. Multimodal document contains both text and images. The framework has three steps: In first step, a semantic model is constructed using knowledge representation based on objects (concepts) organized by ontology. In second step, informational content (concepts) is rated based on information density metric. The metric determines the relevance of concepts based on completeness of attributes, the number of relationships with other concepts and the number of expressions showing the occurrence of concept in the current document. In third step, the important concepts are expressed as sentences. The expressions observed by the parser are stored in a semantic model for expressing concepts and relationship. An important advantage of this framework is that it produces abstract summary, whose coverage is excellent because it includes salient textual and graphical content from the entire document. The limitation of this framework is that it is manually evaluated by humans. An automatic evaluation of the framework is desirable.

3.2.2. 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. A framework proposed in [6] for abstractive summarization took place in the context of Text Analysis Conference (TAC) 2010 for multi-document summarization of news. The framework consists of following modules: Information Item retrieval, sentence generation, sentence selection and summary generation. In Information Item (INIT) retrieval, first syntactic analysis of text is done with parser and the verb's subject and object are extracted. So, an INIT is defined as a dated and located subject–verb–object triple. In sentence generation module, a sentence is directly generated from INIT using a language generator, the NLG realizer Simple NLG [13].Sentence selection module ranks the sentences generated from INIT based on their average Document Frequency (DF) score. Finally, a summary generation step account for the planning stage and include dates and locations for the highly ranked generated sentences. The major strength of this approach is that it produces short, coherent, information rich and less redundant summary. This approach has several limitations. First, many candidate information items are rejected due to the difficulty of creating meaningful and grammatical sentences from them. Secondly, linguistic quality of summaries is very low due to incorrect parses.

3.2.3. 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 abstractive approach proposed by [14] consists of three phases as shown in figure 1. The first Phase represents the input document semantically using Rich Semantic Graph (RSG). In RSG, the verbs and nouns of the input document are represented as graph nodes along with edges corresponding to semantic and topological relations between them. The second phase reduces the generated rich semantic graph of the source document to more reduced graph using some heuristic rules. Finally, the third Phase generates the abstractive summary from the reduced rich semantic graph. This phase accepts a semantic representation in the form of RSG and generates the summarized text. A noteworthy strength of this method is that it produces concise, coherent and less redundant and grammatically correct sentences. However this method is limited to single document abstractive summarization.

4. CONCLUSION

Automatic text summarization is an old challenge but the research direction is leaning from extractive summarization to abstractive summarization. Abstractive summary methods produces coherent, cohesive, information rich and less redundant summary. Because of the complexity of natural language processing, abstractive text summarization is a challenging area. Therefore, this study examines a review on abstractive summarization methods along with their advantages and disadvantages. The different methods are also studied and compared. It is hoped that this study helps the new researchers to get a better understanding of abstractive text summarization techniques.

REFERENCES

[1] H. P. Luhn, "The automatic creation of literature abstracts," *IBM Journal of Research and Development*, vol. 2, pp. 159-165, 1958.

[2] Suneetha Manne, Zaheer Parvez Shaik Mohd. , Dr. S. Sameen Fatima, "Extraction Based Automatic Text Summarization System with HMM Tagger", Proceedings of the International Conference on Information Systems Design and Intelligent Applications, 2012, Vol. 132, P.P 421-428.
[3] Saranyamol C S, Sindhu L, "A Survey on Automatic Text Summarization", International Journal of Computer Science and Information Technologies, 2014,Vol. 5 Issue 6.
[4] Khan Atif, Salim Naomie, "A review on abstractive summarization Methods**",** Journal of Theoretical and Applied Information Technology, 2014, Vol. 59 No. 1.
[5] Reeve Lawrence H., Han Hyoil, Nagori Saya V., Yang Jonathan C., Schwimmer Tamara A., Brooks Ari D., "Concept Frequency Distribution in Biomedical Text Summarization", ACM 15th Conference on Information and Knowledge Management (CIKM), Arlington, VA, USA,2006.
[6] 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.
[7] 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.P.E.
[8] Pierre-Etienne Genest, Guy Lapalme Rali-Diro, "Framework for Abstractive Summarization Using Textto- Text Generation" Universit´e de Montr´eal P.O. Box 6128, Succ. Centre-Ville Montr´eal, Qu´ebec Canada, H3C 3J7.
[9] Pierre-Etienne Genest, Guy Lapalme www.aclweb.org/anthology/P12-2069.
[10] Meghana Viswanath https://getd.libs.uga.edu
[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. F. Greenbacker, "Towards a framework for abstractive summarization of multimodal documents," *ACL HLT 2011,* p. 75, 2011.
[13] 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.
[14] 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.
[15] Ganesan Kavita, Zhai ChengXiang and Han Jiawei, "Opinosis: A Graph-Based Approach to Abstractive Summarization of Highly Redundant Opinions", Proceedings of the 23rd International Conference on computational Linguistics (Coling 2010), 2010, pages 340–348, Beijing.