A Survey on Methods of Abstractive Text Summarization

N. R. Kasture[1], Neha Yargal [2], Neha Nityanand Singh[3], Neha Kulkarni[4] and Vijay Mathur[5]

[1]Prof. N. R. Kasture, Department of Computer Engineering, Vishwakarma Institute of Information Technology, Pune
Sr.No. 2/3/4, Kondhwa (Bk), Pune-48, India.
nehakasture86@gmail.com
[2]Neha Yargal, Department of Computer Engineering, Vishwakarma Institute of Information Technology, Pune
Sr.No. 2/3/4, Kondhwa (Bk), Pune-48, India.
neha.yargal@gmail.com
[3]Neha Nityanand Singh, Department of Computer Engineering, Vishwakarma Institute of Information Technology, Pune
Sr.No. 2/3/4, Kondhwa (Bk), Pune-48, India.
nehasingh2294@yahoo.com
[4]Neha Kulkarni, Department of Computer Engineering, Vishwakarma Institute of Information Technology, Pune
Sr.No. 2/3/4, Kondhwa (Bk), Pune-48, India.
nehaa15kulkarni@yahoo.co.in
[5]Vijay Mathur, Department of Computer Engineering, Vishwakarma Institute of Information Technology, Pune
Sr.No. 2/3/4, Kondhwa (Bk), Pune-48, India.
vijay.mathur88@gmail.com

ABSTRACT

Text summarization is the process of extracting the important information which gives us the overall idea of the entire document. It is a tedious task for human beings to generate an abstract manually since it requires a rigorous analysis of the document. In order to ease human efforts and to reduce time, automatic summarization techniques prove to be very useful. Text summarization has two techniques viz. Extractive summarization and abstractive summarization. Extractive technique[1] is the one in which we generate a summery by using the relevant sentences in the document as they are, whereas in abstractive summarization technique[2] we form the sentences on our own and then combine these sentences to form an abstract. In this paper, an overall idea about the extractive methods is presented and we focus on the abstractive summarization methods.

Keywords – Text Summarization, Extractive summarization, Abstractive Summarization.

1. INTRODUCTION

The need for automatic summarization increases as the amount of textual information increases. A lot of information is available on internet but to sort out the required information is a tedious job. The need for technologies that can do all the sorting and quickly identify the relevant information on its own therefore plays an important role.

Text summarization is a technique which can automatically generate the desired and relevant information from a huge amount of information. The goal of automatic summarization is to form a shorter version of the source document by preserving its meaning and information content. Summarization can be broadly classified into two categories- extractive summarization techniques and abstractive summarization techniques.

Extractive summarization [1] includes selecting important information, paragraphs etc. from a document and combining it to form a new paragraph called as summery. The choice of the sentences depends upon statistical and linguistic features of the sentences. Extractive summaries are formulated by weighting the sentences as a function of high frequency words. Here, the most frequently occurring or the most favourably positioned text is considered to be the most important.

The methods used for determining the weights of the sentences are:

Cue method, location method and title method.

Following were the features used for determining the same:-

Fixed phrase feature, paragraph feature, thematic word feature, uppercase word feature, sentence length cut off feature etc.

Following are the drawbacks of the extractive features:-

-These sentences tend to be longer than average for most of the times. The problem with this is that the parts of the sentences which are not necessary to form the summery also get included thus wasting the space and increasing its length unnecessarily.

-The important and relevant information is usually spread out throughout the document and the extractive techniques are unable to combine all of these unless increasing the size of the summery.

-Also, when the sentences are picked up as they are the pronouns often tend to lose their references thus creating a confusion to trace the meaning.

- If there is a confliction in the information, it may not be presented accurately.

In order to overcome these problems abstractive summarization techniques can be used.

Abstractive summarization [2] includes understanding the main concepts and relevant information of the main text and then expressing that information in short and clear format.

Abstractive summarization techniques can again be classified into two categories- structured based and semantic based methods.

Structured based approaches determines the most important information through documents by using templates, extraction rules and other structures such as tree, ontology etc.

2. RELATED WORK

2.1. Structured Based Abstractive Summarization Methods

2.1.1. Rule Based Method

The rule based method [4] 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.

2.1.2. 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.

2.1.3. Tree Based Method

In this approach, the pre-processing is done of similar sentences using shallow parser [5]. 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.

2.2. Semantic Based Abstractive Summarization

2.2.1. Multimodal Semantic Model

Multimodal semantic model captures the concepts and form the relation among these concepts [6]. These selected concepts are expressed in the form of sentences. This model accepts text document as well as image document.

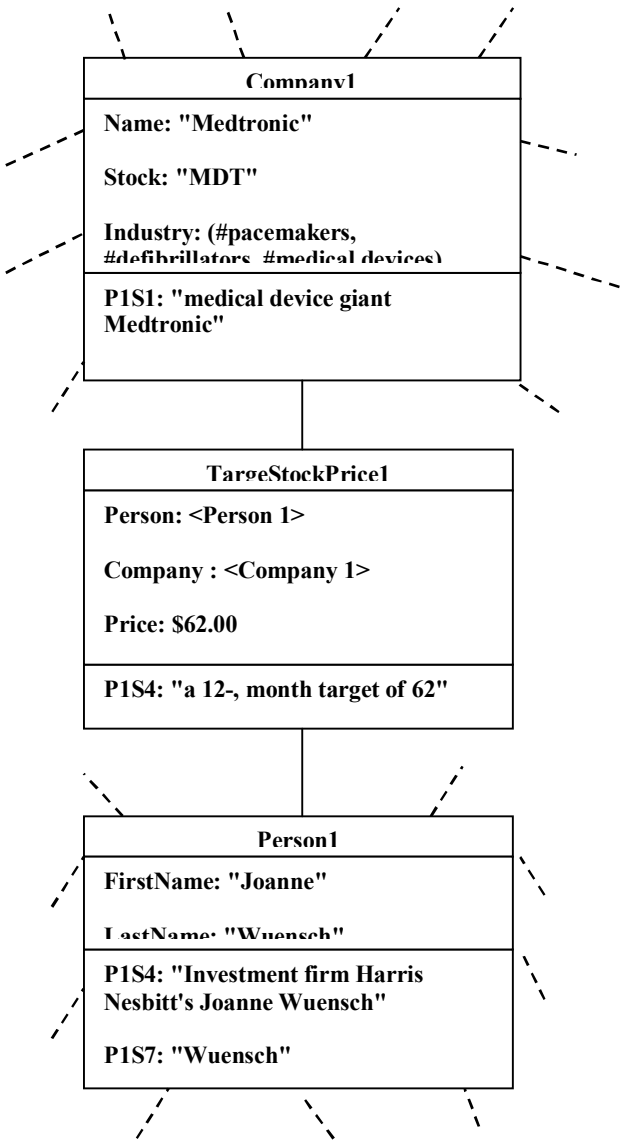

Figure 1: Most Important higher rated concepts included in summary

Multimodal consist of three phases –

[I]. Semantic Modal

Concepts are nothing but words which represents important information. Concepts are constructed using knowledge representation based on objects. Nodes represent concepts and links between these concepts represent relationship between them. Using this semantic models are constructed as shown in Figure 1

[II]. Rated Concepts

Concepts are rated using information density (ID) matrix. This matrix is used to evaluate pertinence of concepts.

Factors for determining the relevance –

Completeness of attributes- it is nothing but the ratio of filled attributes in the semantic model to the total number of attributes in semantic. This gives sentences which contain more information.

One concept is connected with other concepts and these relations are counted. Keeping a track of this count helps us to know how important this concept is.

[III]. Sentence Generation

Once the concepts are rated using ID matrix the next step is to generate sentences using parsing techniques.

2.2.2. Information item based method

In this method, instead of generating abstract from sentences of the input file, it is generated from abstract representation of the input file. The abstract representation is nothing but an information item which is the smallest element of information in a text. The framework [8] used in his method was proposed in the context of Text Analysis Conference (TAC) 2010 for multi-document summarization of news. The modules of this framework are: Information item retrieval, sentence generation, sentence selection and summary generation.

In Information Item (INIT) retrieval phase, subject-verb-object triples are formed by syntactical analysis of text done with the help of parser. While syntactical analysis, verb's subject and object are extracted. In sentence generation phase, the sentences are generated using a language generator. In the next phase i.e. sentence selection phase, raking of each sentence is done on the basis of the average document frequency (DF)

score. At last in summary generation phase, highly ranked sentences are arranged and abstract is generated with proper planning.

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.

2.2.3. Semantic Graph Based Methods

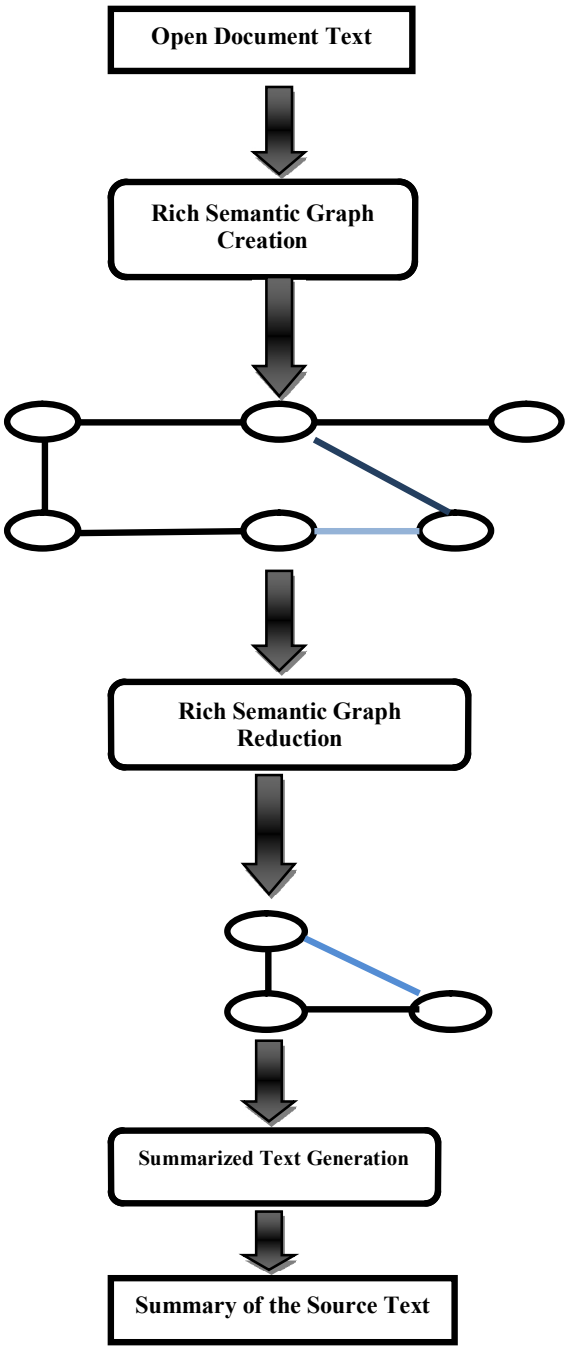

Figure 2: Semantic Graph Reduction

The main objective of this method is generating a summery by creating a semantic graph called rich semantic graph (RSG) [3]. As shown in Figure 2

The semantic graph approach consists of three phases:-

-The first phase represents input document using rich semantic graph (RSG). In RSG, 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.

3. CONCLUSION

In text summarization, the greatest challenge is to retrieve relevant information from given structural sources including web pages, any document, and database. An effective summery must be produced by text summarization techniques using less time and less redundancy. From the above studied methods, the advantages and limitations of each method is written below. The summary generated by the Rule based technique is of high information density but it is very tedious work because all the rules and patterns are written manually. In the method of Ontology, handling of uncertain data is possible which is not possible in simple domain ontology. Problem with this method is that only domain experts can define the ontology of the domain which is time consuming. In the Tree based technique, the quality of summary gets improved because of the use of language generator. Only problem with this method is that the main context of the sentences gets rejected while capturing the intersection of phrases. The Multimodal semantic model method produces abstract summary in which it includes textual data as well as graphical data and hence, gives excellent result. Problem with this method is that evaluation is to be done manually. In the Information item based method, the selection of useful information is done. On the basis of selected information item the sentences and summaries are generated. This approach gives a small, coherent and information rich summary.

Problem with this method is that sometimes useful information items gets rejected while the construction of meaningful and grammatically correct sentences which reduces the linguistic quality of summary. The Semantic graph method, Sentences formed are less redundant as well as grammatically correct. But this method is limited to only single document.

Though the technique of automatic summarization is an old challenge, the experts are nowadays getting more inclined towards abstractive summarization techniques rather than extractive summarization techniques. This is because, abstractive summarization methods produce more coherent, less redundant and information rich summery. Generating abstract using abstractive summarization methods is a difficult task since it requires more semantic and linguistic analysis. Due to the above reasons the study of abstractive summarization techniques proves to be more useful.

REFERENCES

[1] Vishal Gupta, Gurpreet Singh Lehal, "A Survey of text summarization of extractive techniques" University institute of engineering and Technology, Computer Science & Engineering, Punjab University, Chandigarh, India, E-mail: vishal@pu.ac.in, gslehal@yahoo.com.

[2] Atif Khan, Naomie Salim**, "A** Review on Abstractive Summarization Methods**"** Faculty of Computing, Universiti Teknologi Malaysia, 81310, Skudai, Johor, Malaysia, E-mail: katif2@utm.my, naomie@utm.my.

[3] Ibrahim F. Moawad, Mostafa Aref, "Semantic Graph Reduction Approach for Abstractive Text Summarization" Information Systems Dept. Faculty of Computer and Information Sciences, Ain shams University Cairo, Egypt.

[4] Pierre-Etienne Genest, Guy Lapalme Rali-Diro,"Fully Abstractive Approach to Guided Summarization" Universit´e de Montr´eal P.O. Box 6128, Succ. Centre-Ville Montr´eal, Qu´ebec Canada, H3C 3J7.

[5] Pierre-Etienne Genest, Guy Lapalme Rali-Diro, "Framework for Abstractive Summarization Using Text-to-Text Generation" Universit´e de Montr´eal P.O. Box 6128, Succ. Centre-Ville Montr´eal, Qu´ebec Canada, H3C 3J7.

[6] Charles F.Greenbacker,"Towards a Framework for Abstractive Summarization of Multimodal Documents" Dept of Computer and Information Sciences Universitiy of Delaware Newark, U.S.A.

[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] 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.

