ISSN:0975-9646

A Survey on Automatic Text Summarization

Saranyamol C S[#1],Sindhu L [#2]

Dept.of Computer Science and Engineering ,
College of Engineering Poonjar
Cochin University of science and Technology,Kerala,India

***Abstract*-With the proliferation of online information, text summarization has become essential to provide enhanced mechanisms to perceive and present effective textual information. It is very difficult for human beings to manually summarize large documents of text. Automatic text summarization has become an important and timely tool for assisting and interpreting text information. Text summarization is the process of automatically creating a condensed form of a given document retaining its information content. This survey paper describes different approaches to the automatic text summarization process and making an analysis of different methods.**

Keywords-Text summarization, Natural language processing, Extractive summary, Abstractive summary

1. INTRODUCTION

Text summarization is a way to condense the large amount of information into a concise form by the process of selection of important information and discarding unimportant and redundant information. With the amount of textual information present in the World Wide Web (www), the area of Automatic Text Summarization (ATS) is becoming very important in the field of information retrieval. The process of condensing a source text in to a shorter version preserving its information content is called summarization. Automated summarization tools can help people to grasp main concepts of information sources in a short time.

Text summarization approaches can be broadly divided into two groups: extractive summarization and abstractive summarization. Extractive summarization extracts salient sentences or phrases from the source documents and group them to produce a summary without changing the source text. Usually, sentences are in the same order as in the original document text. However, abstractive summarization consists of understanding the source text by using linguistic method to interpret and examine the text. The abstractive summarization aims to produce a generalized summary, conveying in information in a concise way, and usually requires advanced language generation and compression techniques[1]. .Early work in summarization started with single document summarization. Single document summarization produces summary of one document. As research proceeded, and due to large amount of information on web, multi document summarization emerged. Multi document summarization produces summaries from many source documents on the same topic or same event. The automatic summarization of text is a well- known task in the field of natural language processing (NLP). Significant achievements in text summarization have been obtained using sentence extraction and statistical analysis. True abstractive summarization is a dream of researchers [2]. 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 .

One of examples of document summarization is in the field of legal area. The legal experts perform difficult and responsible work. Their resources are sparse and expensive both in time and expertise levels. Thus, the system for concise summarization is necessary in order for experts to be able to effectively and in short time find compressed and restated content of relevant judicial documents, including laws and their proposals, relevant court decisions or tribunal process summarizations etc. In the medical branch, there is often overload of information and it is requirement in many cases for the medical personal to find relevant information about patient's conditions every time. This involves crawling of many documents and patient's record in order to gain necessary information. In this area the text summarization specifically adjusted to medical domain is of considerable use, saving time resources and optimizing availability of medical experts.

On the internet, there are many examples of the summarizations used. For instance, news portals like Microsoft News2,Google1 or Columbia Newsblaster3 are relying of such techniques in order to provide short news summaries to their visitors. There are also service providing blog summarization and aggregation4 and opinion survey systems.

Further, there is also application of document summarization for PDA devices with small screen, where the only limited screen size and time are available for users to read. For the businesses it is also important to have available summarizations of meetings coupled possibly with speech recognition systems, to provide "meeting minutes" in short time and without using excessive human and other resources. For the handicapped people, the text summarization systems are also of great help. They can save much time for readers of such documents using speech synthesis technologies, and in order to be able to recognize and separate important and less important content according to their interests.

2. TECHNIQUES USED FOR TEXT SUMMARIZATION

Text summarization can be broadly divided into extractive and abstractive. Some of the methods are discussed below.

2.1 Extractive Summarization Techniques

2.1.1 *Term Frequency-Inverse Document Frequency (TF-IDF) method:*

Sentence frequency is defined as the number of sentences in the document that contain that term. Then this sentence vectors are scored by similarity to the query and the highest scoring sentences are picked to be part of the summary.

2.1.2 *Cluster based method*

Documents are usually written such that they address different topics one after the other in an organized manner. They are normally broken up explicitly or implicitly into sections. It is intuitive to think that summaries should address different "themes" appearing in the documents. Some summarizers incorporate this aspect through clustering. If the document collection for which summary is being produced is of totally different topics, document clustering becomes almost essential to generate a meaningful summary. Sentence selection is based on similarity of the sentences to the theme of the cluster Ci .The next factor that is considered for sentence selection is the location of the sentence in the document (Li). the closer to the beginning a sentence appears, the higher its weight age for inclusion in summary. The last factor that increases the score of a sentence is its similarity to the first sentence in the document to which it belongs (Fi).

The overall score (Si) of a sentence i is a weighted sum of the above three factors:

Si =W1 *Ci + W2 *Fi+ W3 *Li

2.1.3 *Graph theoretic approach*

Graph theoretic representation of passages provides a method of identification of themes. After the common preprocessing steps, namely, stemming and stop word removal sentences in the documents are represented as nodes in an undirected graph. There is a node for every sentence. Sentences are connected with an edge if the two sentences share some common words. For query-specific summaries, sentences may be selected only from the pertinent sub graph, while for generic summaries, representative sentences may be chosen from each of the sub-graphs. The nodes with high cardinality are the important sentences in the partition, and hence gets higher preference to be included in the summary.

2.1.4 *Machine Learning approach*

Given a set of training document and their extractive summaries, the summarization process is modeled as a classification problem: sentences are classified as summary sentences and non-summary sentences based on the features that they possess. The classification probabilities are learnt statistically using Bayes' rule:

$P(s \in S \mid F1, F2, ..., FN) = P(F1, F2, ..., FN \mid s \in S) * P(s \in S) / P(F1, F2, ..., FN)$

where S is a sentence from the document collection, F1, F2…FN are features which are used for classification. S is the summary to be generated, and $P(s \in S \mid F1, F2, ..., FN)$ is the probability that sentence s will be chosen to form the summary given that it possesses features F1,F2…FN.

2.1.5 LSA Method

It gets this name Latent Semantic Analysis because Singular Value Decomposition applied to document- word matrices that are semantically related to each other, even though they do not share common words. Words that usually occur in related contexts are related in the same singular space. This method can be applied to extract the content-sentences and topic-words from documents. The advantage of using LSA vectors for summarization rather than the word vectors is that conceptual (or semantic) relations as represented in human brain are automatically captured in the LSA, while using word vectors without the LSA transformation requires design of explicit methods to derive conceptual relations.

2.1.6 *Text summarization with neural networks*

This method involves training the neural networks to learn the types of sentences that should be included in the summary. This is done by the human reader. The neural network learns the patterns that are inherent in sentences and that should be included in the summary and those that should not be included. It uses three-layered Feed forward neural network, which is proven to be a universal function approximator.

2.1.7 *Automatic text summarization based on fuzzy logic*

This method considers each characteristic of a text such as similarity to little ,sentence length and similarity to key word etc. as the input of the fuzzy system. Then, it is entering all the rules needed for summarization in the knowledge base of the system. After that a value from zero to one is obtained for each sentence in the output based on sentence characteristics and the available rules in the knowledge base. Then the obtained value in the output determines the degree of the importance of the sentence in the final summary.

2.1.8 *Query based extractive text summarization*

In query based text summarization system, the sentences in a given document are scored based on the frequency counts of terms . The sentence that containing the query phrases are given higher scores than the ones containing single query words. Then, the sentences with highest scores are incorporated into the output summary together with their structural context. Portions of text may be extracted from different sections or subsections. The resulting summary is the union of such extracts.

2.2 Abstractive Summarization Techniques

Summarization using abstractive techniques are broadly classified into two categories: Structured based approach and Semantic based approach.

2.2.1 Structured Based Approach

Structured based approach encodes most important information from the document through cognitive schemas

such as templates, extraction rules and other structures such as tree, ontology, lead and body phrase structure.

2.2.1.1 Tree based method

This technique uses a dependency tree to represent the text/contents of a document. Different algorithms are used for content selection for summary such as theme intersection algorithm or algorithm that uses local alignment across pair of parsed sentences. The technique uses either a language generator or an algorithm for generation of summary. The limitation of this approach is that it lacks a complete model which would include an abstract representation for content selection.

2.2.1.2 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. The templates are filled with important text snippets extracted by the Information Extraction systems. A significant advantage of this approach is that the generated summary is highly coherent because it relies on relevant information identified by IE system. This approach works only if the summary sentences are already present in the source documents. It cannot handle the task if multi document summarization requires information about similarities and differences across multiple documents.

2.2.1.3. Ontology based method

Many researchers have made effort to use ontology(knowledge base) to improve the process of summarization. Most documents on the web are domain related because they discuss the same topic or event. Each domain has its own knowledge structure and that can be better represented by ontology. The benefit of this approach is that it exploits fuzzy ontology to handle uncertain data that simple domain ontology cannot. This approach is limited to Chinese news only.

2.2.1.4 Lead and body phrase method

This method is based on the operations of phrases (insertion and substitution) that have same syntactic head chunk in the lead and body sentences in order to rewrite the lead sentence. The potential benefit of this method is that it found semantically appropriate revisions for revising a lead sentence. This method has some weaknesses. First, Parsing errors degrade sentential completeness such as grammaticality and repetition. Secondly, it focuses on rewriting techniques, and lacks a complete model which would include an abstract representation for content selection.

2.2.1.5 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 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.

2.2.2 Semantic Based Approach

In Semantic based method, semantic representation of document is used to feed into natural language generation (NLG) system. This method focus on identifying noun phrases and verb phrases by processing linguistic data

2.2.2.1 Multimodal semantic model

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

2.2.2.2 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. 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.

2.2.2.3 Semantic Graph Based Method

This methodis used to summarize a document by creating a semantic graph called Rich Semantic Graph (RSG) for the original document, reducing the generated semantic graph. After that it generating the final abstractive summary from the reduced semantic graph. 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.

3.RELATED WORK

Naresh Kumar Nagwani et al[3] proposed a method for text summarization based on frequent terms and semantic similarity. In this method the system is divided into three parts: an input text document, a summarizer algorithm and the summarized text document. The summarizer algorithm is having two steps: text pre processing module and frequent term generation module along with the semantically similar term and sentence filtering module for summarization.The overall methodology of semantic similarity based single document summarization can be

expressed in terms of an algorithm. The algorithm takes two input parameters: the input text document and number of frequent terms. As the output it generates a summarized text document along with the two measures compression ratio and retention ratio. Compression ratio measures how much shorter the summary is than the original document. Retention ratio measures how much of the central information is retained. This algorithm is applied here for single document, it can be applied for multi document.

Divya S et al[4] proposed a method for text summarization based on sentence clustering and ranking. In this method it identifies common information among multiple related news documents and fuse it into a coherent text to produce an abstract of the news events. Here it proposes a statistical approach for sentence clustering based on similarity of sentences followed by sentence ranking based on eigen vectors. The important stages involved in this method are retrieving related documents, sentence clustering, sentence ranking and linking sentences. User query is processed by stemming and stop word elimination and then identifying keyword. Keywords are using for retrieving relevant documents. A customized web crawler program is used for retrieving news articles from news providers and is stored as a repository and later indexed. Sentence clustering component takes care of redundant information in related news articles. Similar sentences are grouped together to form a cluster. Similarity between sentences is performed with statistical approach known as term frequency inverse document frequency. The retrieved documents are tokenized into words, preprocessed by stemming and stop word elimination.

Sentence ranking is used to find the important sentence. For that a sentence similarity matrix is created. Similarity of two sentence i and j in the matrix is measured by counting the number of key terms similar in sentence I and j. Then the vector corresponding to the characteristic value(Eigen value) of the similarity matrix will give the most correlated sentence in cluster. After identifying the important sentence they are arranged in a meaningful order by sentence linking component.

Dr.P C Reghu Raj et al[5] proposed text summarization by a method of Eigen vector based approach. Here sentence ranking is done by using Eigen vectors. Sentence similarity matrix is generated for each cluster and Eigen vector for the matrix is calculated. This approach is capable of handling redundant information in such a way that users will get all related news events as a single news summary. Sentence ranking is based on Eigen vector centrality and is a measure of the importance of a node in a network.

A.R.Kulkarni et al[6] proposed a method for summarization by lexical cohesion and correlation of sentences. Lexical chains are created by wordnet. The score of each lexical chain is calculated based on keyword strength, term frequency inverse document frequency. The concept of using lexical chain helps to analyze the document semantically and the concept correlation of sentences helps to consider the relation of sentence as the preceding or succeeding sentence. This improves the quality of summary. Here the document is preprocessd first.. Preprocessing includes Segmentation Tokenization POS(part of speech tagging) at lexical level and Stemming. Lexical chain computing algorithm is used for finding lexical chain. Evaluation is based on two parameters known as precision and recall.

$$\text{Precision} = \frac{\{\text{Retrieved sentences}\}-\{\text{Relevant sentences}\}}{\{\text{Retrieved Sentences}\}}$$

$$\text{Recall} = \frac{\{\text{Retrieved sentences}\}-\{\text{Relevant sentences}\}}{\{\text{relevant sentences}\}}$$

Sreejith C et al[7] describes an approach for information box article generation via the paragraph ranking method. Paragraph ranking is performed with the help of the term frequency measuring and vector space modeling methods. In this paper, a method to automate the process of text document to create box article is proposed based on the term frequency within the document at different levels paragraph and sentence. The similarity between the paragraphs and sentence within the paragraphs are considered using the vector space model. For the generation of box article it uses three modules: Pre-processing module, Paragraph ranking module and Filtering module . Pre-processing of the text document is done to obtain a structured representation of the original text.

Paragraphs in the document are ranked according to their significant relevance to the document. The similarity of two paragraphs i and j in the matrix is measured by the term frequency. In vector model, the frequency of a term , in the document , is referred to as the term frequency factor.

The tf factor is given by, $tf = \frac{freq\ i,j}{max\ lfreq\ l,j}$

Where,$freqi,j$, is the frequency of term ki in the document dj.

In the filtering module paragraph with high score is selected. Paragraph with highest score is included as the news box item, which retains the overall semantic meaning of the original text document.

3. CONCLUSION

The automatic summarization of text is a well- known task in the field of natural language processing (NLP). Text summarization cane be broadly divided into two categories: extractive and abstractive summarization . This paper deals with the techniques used in both the methods.

The importance of sentences is decided based on statistical and linguistic features of sentences. Without the use of NLP, the generated summary may suffer from lack of cohesion and semantics. If texts containing multiple topics, the generated summary may not be balanced.

Automatic text summarization is an old challenge but now the research direction is leaning from extractive text summarization to abstractive text summarization. In abstractive summarization methods it produces cohesive, highly coherent,information rich and less redundant summary. Abstractive text summarization is a challenging area because of the complexity of natural language processing.

REFERENCES

[1] D. Das and A. F. Martins, "A survey on automatic text summarization," Literature Survey for the Language and Statistics II course at CMU, vol. 4, pp. 192-195, 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] Naresh Kumar Nagwani ,Dr. Shrish Verma Associate "A Frequent Term and Semantic Similarity based Single Document Text Summarization Algorithm" , International Journal of Computer Applications (0975 – 8887) Volume 17– No.2, March 2011

[4] Divya S., Dr. P. C. Reghuraj Department of Computer Science and Engineering Govt. Engg. College Sreekrishnapuram "News Summarization Based on Sentence Clustering and Sentence Ranking"

[5] Divya S., Dr. P. C. Reghuraj Department of Computer Science and Engineering Govt. Engg. College Sreekrishnapuram "Eigenvector Based Approach for Sentence Ranking in News Summarization", International Journal of Computational Linguistics and Natural Language Processing Vol 3 Issue 3 March 2014

[6] Sreejith C, Sruthimol M P *2, P C Reghuraj @3 M.Tech, Department of Computer Science and Engineering, Government Engineering College, Palakkad "Box Item Generation from News Articles Based Paragraph Ranking using Vector Space Model", International Journal of Scientific Research in Computer Science Applications and Management Studies IJSRCSAMS Volume 3, Issue 2 (March 2014)

[7] A.R.Kulkarni1, S.S.Apte2," An Automatic Text Summarization Using Lexical Cohesion And Correlation Of Sentences " , IJRET: International Journal of Research in Engineering and Technology eISSN: 2319-1163 | pISSN: 2321-7308