An Overview on Document Summarization Techniques

Archana AB, Sunitha. C

Computer Science Department, Calicut University, Computer Science Department Calicut University

Abstract – The increase in the performance and fast accessing of web resources has made a new challenge of browsing among huge data on Internet. Since digitally stored information is more and more available, users need suitable tools able to select, filter, and extract only relevant information. Text summarization can be classified into two approaches: extraction and abstraction. This paper focuses on extraction approach. The goal of text summarization based on extraction approach is sentence selection. One of the methods to obtain the suitable sentences is to assign some numerical measure of a sentence for the summary called sentence weighting and then select the best ones. Therefore, text summarization techniques are currently adopted in several fields of information retrieval and filtering, such as, information extraction, text mining, document classification, recommender systems, and contextual advertising. A summary text is a derivative of a source text condensed by selection and/or generalization on important content. Query-focused summaries enable users to find more relevant documents more accurately, with less need to consult the full text of the document. Extractive summarization methods try to find out the most important topics of an input document and select sentences that are related to these chosen concepts to create the summary. This paper is a Comparative study of four techniques used for extractive summarization, namely, Neural Network, Graph Theoretic, Fuzzy based method and Cluster based method

Index Terms—Cluster, Cardinality, pruning, Rule Base.

I. INTRODUCTION

In the sixties, a large amount of scientific papers and books have been digitally stored. However, the storage media to store such a large database was very expensive. Therefore the concept of automatic shortening of texts was introduced to store the information about papers and books in limited storage space. Now, due to advancement in technology, the storage media are no longer expensive and bulk of information can be fit into the large databases these days[1]. But due to increased use of the Internet, and large amount of information available on the web, there is a need to represent each document by its summary to save time and effort for searching the correct information.

In summary, the following are the important reasons in support of automatic text summarization:

1) A summary or abstract saves reading time
2) A summary or an abstract facilitate document selection and literature searches.
3) It improves document indexing efficiency
4) Machine generated summary is free from bias
5) Customized summaries can be useful in question an-swering systems where they provide personalized information.
6) The use of automatic or semi-automatic summarization by commercial abstract services may allow them to scale the number of published texts they can evaluate.

Automatic document summarization is extremely helpful in tackling the information overload problems. It is the technique to identify the most important pieces of information from the document, omitting irrelevant information and minimizing details to generate a compact coherent summary document. There are different types of summarization approaches depending on what the summarization method focuses on to make the summary of the text [2].

1) **Abstract vs. Extract summary** - Abstraction is the process of paraphrasing sections of the source document whereas extraction is the process of picking subset of sentences from the source document and presents them to user in form of summary that provides an overall sense of the documents content.

2) **Generic vs. Query-based summary** - Generic summary do not target to any particular group. It addresses broad community of readers while Query or topic focused queries are tailored to the specific needs of
an individual or a particular group and represent particular topic.

3) **Single vs. Multi-document summary** - Single document summary provide the most relevant information contained in single document to the user that helps the user in deciding whether the document is related to the topic of interest or not whereas multi-document summary helps to identify redundancy across documents and compute the summary of a set of related documents of a corpus such that they cover the major details of the events in the documents, taking into account some of the major issues: the need to carefully eliminate redundant information from multiple documents and achieve high compression ratios; information about document and passage similarities, and weighting different passages accordingly; the importance of temporal information; co-reference among entities and facts occurring across documents.

4) **Indicative vs. informative**: An indicative summary provides merely an indication of the principal subject matter or domain of the input text(s) without including its contents. After reading an informative summary, one can explain what the input text was about, but not necessarily what was contained in it. An informative summary reflects (some of) the content, and allows one to describe (parts of) what was in the input text.

5) **Background vs. just-the-news**: A background summary assumes the reader’s prior knowledge of the general setting of the input text(s) content is poor, and hence includes explanatory material, such as circumstances of place, time, and actors. A just-the-news summary contains just the new or principal themes, assuming that the reader knows enough background to interpret them in context.

### A. Approaches To Text Summarization

Each approach has got a unique way of dealing with the text summarization problem. Graph theoretic approach represents all sentences as nodes of a graph to extract the important sentences in a document. In the Neural Network based approach, a neural network is trained to pick out the important sentences. Inference engine and rule base are the main components of the Fuzzy based system. In cluster based system sentences are converted to triplets and they are then grouped together, based on the information they carry. In this section, we are presenting an outline of the four different approaches.

#### 1. **Graph Theoretic Approach**

Graph theoretic representation of passages provides a method of identification of themes. After the common preprocessing steps, namely, stop word removal and stemming, sentences in the documents are represented as nodes in an undirected graph.

There is a node for every sentence. Two sentences are connected with an edge if the two sentences share some common ords, or in other words, their (cosine, or such) similarity is above some threshold. This representation yields two results: The partitions contained in the graph (that is those sub-graphs that are unconnected to the other sub-graphs), form distinct topics covered in the documents [3]. This allows a choice coverage in the summary. 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 second result yielded by the graph-theoretic method is the identification of the important sentences in the document. The nodes with high cardinality (number of edges connected to that node), are the important sentences in the partition, and hence carry higher preference to be included in the summary.

Figure 1 shows an example graph for a document. It can be seen that there are about 3-4 topics in the document; the nodes that are encircled can be seen to be informative sentences in the document, since they share information with many other sentences in the document. The graph theoretic method may also be adapted easily for visualization of inter and intra document similarity.

---

**Graph Theoretic Approach**

**II. TEXT SUMMARIZATION WITH NEURAL NETWORKS**

In this method, each document is converted into a list of sentences. Each sentence is represented as a vector \([f_1, f_2, \ldots, f_7]\), composed of 7 features. Take a machine learning approach that uses artificial neural networks to produce summaries of arbitrary length news articles. News articles have a highly structured document form; important ideas can be obtained from the text simply by selecting sentences based on their attributes and locations in the article. A neural network is trained on a corpus of articles. The neural network is then modified, through feature fusion, to produce a summary of highly ranked sentences of the article. Through feature fusion, the network discovers the importance (and unimportance) of various features used to determine the summary-worthiness of each sentence [4].

**Seven Features of a Document**

1) **f1 Paragraph follows title**
2) f2 Paragraph location in document
3) f3 Sentence location in paragraph
4) f4 First sentence in paragraph
5) f5 Sentence length
6) f6 Number of thematic words in the sentence
7) f7 Number of title words in the sentence

Feature f5, sentence length, is useful for filtering out short sentences such as datelines and author names commonly found in news articles. It is also anticipated that short sentences are unlikely to be included in summaries. Feature f6, the number of thematic words, indicates the number of thematic words in the sentence, relative to the maximum possible. It is obtained as follows: from each document, remove all prepositions, and reduce the remaining words to their morphological roots. The resultant content words in the document are counted for occurrence. The top 10 most frequent content words are considered as thematic words. This feature determines the ratio of thematic words to content words in a sentence. This feature is expected to be important because terms that occur frequently in a document are probably related to its topic. Therefore, a high occurrence of thematic words in salient sentences is expected. Finally, feature f7 indicates the number of title words in the sentence, relative to the maximum possible. It is obtained by counting the number of matches between the content words in a sentence, and the words in the title. This value is then normalized by the maximum number of matches. This feature is expected to be important because the salience of a sentence may be affected by the number of words in the sentence also appearing in the title. These features may be changed or new features may be added. The selection of features plays an important role in determining the type of sentences that will be selected as part of the summary and, therefore, would influence the performance of the neural network. There are mainly two different phases in this system.

a) Neural Network Training: The first phase of the process involves training the neural networks to learn the types of sentences that should be included in the summary. This is accomplished by training the network with sentences in several test paragraphs where each sentence is identified as to whether it should be included in the summary or not. This is done by a human reader. The neural network learns the patterns inherent in sentences that should be included in the summary and those that should not be included. A three-layered feed forward neural network is used, which has been proven to be a universal function approximator. It can discover the patterns and approximate the inherent function of any data to an accuracy of 100 percentage, as long as there are no contradictions in the data set. The neural network consists of seven input-layer neurons, six hidden-layer neurons, and one output-layer neuron. A conjugate gradient method where the energy function is a combination of error function and a penalty function is used. The goal of training is to search for the global minima of the energy function. The addition of the penalty function drives the associated weights of unnecessary connections to very small values while strengthening the rest of the connections. Therefore, the unnecessary connections and neurons can be pruned without affecting the performance of the network.

Once the network has learned the features that must exist in summary sentences, there is a need to discover the trends and relationships among the features that are inherent in the majority of sentences. This is accomplished by the feature fusion phase, which consists of two steps: 1) eliminating uncommon features; and 2) collapsing the effects of common features. The connections having very small weights after training can be pruned without affecting the performance of the network. As a result, any input or hidden layer neuron having no emanating connections can be safely removed from the network. In addition, any hidden layer neuron having no abutting connections can be removed. This corresponds to eliminating uncommon features from the network.
2) **Sentence Selection:** Once the network has been trained, pruned, and generalized, it can be used as a tool to filter sentences in any paragraph and determine whether each sentence should be included in the summary or not. This phase is accomplished by providing control parameters for the radius and frequency of hidden layer activation clusters to select highly ranked sentences. The sentence ranking is directly proportional to cluster frequency and inversely proportional to cluster radius. Only sentences that satisfy the required cluster boundary and frequency are selected as high-ranking summary sentences.

B. **Text summarization based on fuzzy logic**

Fuzzy logic system design usually implicates selecting fuzzy rules and membership function. The selection of fuzzy rules and membership functions directly affect the performance of the fuzzy logic system.

The fuzzy logic system consists of four components: fuzzifier, inference engine, defuzzifier, and the fuzzy knowledge base. In the fuzzifier, crisp inputs are translated into linguistic values using a membership function to be used to the input linguistic variables. After fuzzification, the inference engine refers to the rule base containing fuzzy IF-THEN rules to derive the linguistic values. In the last step, the output linguistic variables from the inference are converted to the final crisp values by the defuzzifier using membership function for representing the final sentence score \[5\].

Inference engine, the most important part in this procedure is the definition of fuzzy IF-THEN rules. The important sentences are extracted from these rules according to our features criteria. Sample of IF-THEN rules shows as the following rule.

IF (NoWordInTitle is VH) and (SentenceLength is H) and (TermFreq is VH) and (SentencePosition is H) and (SentenceSimilarity is VH) and (NoProperNoun is H) and (NoThematicWord is VH) and (NumericalData is H) THEN (Sentence is important) Likewise, the last step in fuzzy logic system is the defuzzification. The output membership function which is divided into three membership functions: Output Unimportant, Average, and Important is used to convert the fuzzy results from the inference engine into a crisp output for the final score of each sentence.

In fuzzy logic method, each sentence of the document is represented by sentence score. Then all document sentences are ranked in a descending order according to their scores. A set of highest score sentences are extracted as document summary based on the compression rate. It has been proven that the extraction of 20 percent of sentences from the source document can be as informative as the full text of a document. Finally, the summary sentences are arranged in the original order.

C. **Text summarization using cluster based method**

In this technique, the semantic nature of a given document captured, expressed in natural language, by a set of triplets (subjects, verbs, objects related to each sentence) and to cluster these ones thus aggregating similar information. In our vision, the triplets statements can be considered as the basic unit in the process of summarization and the more similar the triplets are, the more the information is useless repeated; thus, a summary may be thus constructed using a sequence of sentences related the computed clusters.

1) **Reporter finds the following documents:**

By means of a standard engine,\[6\] our reporter finds the following documents: A Rockwell Commander 112 airplane crashed into the upper floors of the Pirelli Tower in Milan, Italy. Police and ambulances are at the scene. The president, Marcello Pera, just moments ago was informed about the incident in Milan, he said at his afternoon press briefing. It was the second time since the Sept 11 terror attacks on New York and Washington that a plane has struck a high-rise building. Many people were on the streets as they left work for the time of the crash. Ambulances streamed into the area and pedestrians peered upward at the sky. The clock fell to the floor. In Rome, a spokesman for the senate president, Marcello Pera, said the interior minister had informed him that the crash did not appear to be a terror attack.

**Summary of above News Report:**

A Rockwell Commander 112 airplane crashed into the upper floors of the Pirelli Tower in Milan, Italy. Police
and ulances are at the scene. In Rome, a spokesman for the senate president, Marcello Pera, said the interior minister had informed him that the crash did not appear to be a terror attack.

Clusters of Triplets

To automatically summarize such documents, first, reduce each sentence to a list of triplets formed by (subject, verb, object), the verb being reported in the infinitive form. Then clusterize the triplets, using a semantic similarity function between each pair of triplets.

Figure 6 reports the extracted triplets grouped in 3 clusters; underlined triplets represent the three centroids that allow to obtain the above summary.

III. COMPARISON OF ABOVE TECHNIQUES

Graph Theoretic Method gives a better representation of the original document. Sentences having similar context will be grouped together. It gives an idea on how many topics are covered in the document. It can also be used to visualize inter and intra document similarity. Graph-based ranking algorithms do not require training data. They are also easily adaptable to other languages. The disadvantage of Graph Theoretic Method is that the Ranking algorithm does not give a smooth summary. Also a matching word does not guarantee any kind of similarity.

The Advantage of Neural network is that they cope well with noisy data. They are easier to modify for new user communities. Neural Network does not use re-programmed knowledge base. They are user friendly, robust, flexible and gives a clear output. On the downside, an intruder can train the net during its learning phase. Neural networks cannot always explain why they arrived at a particular solution. Moreover, they cannot always guarantee a completely certain solution, arrive at the same solution again with the same input data, or always guarantee the best solution. They are also very sensitive and may not perform well if their training covers too little or too much data.

The Fuzzy system uses a rule base, which means that the volume of knowledge to program is not a major concern. Whether the rule base has 10 rules or 10 000, the engine operation is the same. Evolving an expert system is to add, modify or delete rules. Since the rules are written in plain language, it is easy to identify those to be removed or modified. The system can explain how the summary was generated. The most difficult part of Fuzzy system is building the Rule base.

Since clusters are made after semantic analysis, it is much more effective. A sentence is chosen from each cluster, so it almost sure that the summary will cover all topics in the original document. The disadvantage of this method is that we may get too large or too small clusters. The length of the summary will depend on the number of clusters.

IV. CONCLUSION AND FUTURE WORK

Automated summarization is an important area in NLP (Natural Language Processing) research. It consists of automatically creating a summary of one or more texts. The purpose of extractive document summarization is to automatically select a number of indicative sentences, passages, or paragraphs from the original document according to a target summarization ratio and then sequence them to form a concise summary. Summaries as short as 17 percent of the full text length speed up decision making twice, with no significant degradation in accuracy. In this paper, we have done a comparative study on four different approaches to automatic text summarization. Text summarization approaches based on Neural Network, Graph Theoretic, Fuzzy and Cluster have, to an extent, succeeded in making an effective summary of a document.

Researches on text summarization had started more than 50 years ago, but still there is a long way to go. Both extractive and abstractive methods have been researched. Most summarization techniques are based on extractive methods. Abstractive method is similar to summaries made by humans. Abstractive summarization as of now requires heavy machinery for language generation and is difficult to replicate or extend to broader domains. There is a need to bridge the gap between human-written summaries and machine generated summaries. Next step towards this goal is to take the summary generated from extractive techniques and refine and rephrase them to form an abstractive summary. Linguistics and semantic analysis can definitely open the door for an effective abstractive technique.

V. REFERENCES

[1] U. Hahn and I. Mani,”The challenges of automatic summarization”, IEEE-Computer Publication, 2000, Vol. 33, No. 11, pp. 29-36. in press.

[2] V. Gupta and G.S. Lehal, "A Survey of Text Summarization Extractive Techniques"Journal of
Emerging Technologies in Web Intelligence, Vol. 2, No. 3, August 2010, pp. 258-268.

[3] S.T. Khushboo, R.V. Dharaskar and M.B. Chandak, "Graph based Algorithms for Text Summarization", Third International Conference on Emerging Trends in Engineering and Technology, 2010 pp. 49-53.

[4] L. Suanmali, N. Salim and M.S. Binwahlan, "Fuzzy Logic Based Method for Improving Text Summarization", International Journal of Computer Science and Information Security, 2009, Vol. 2, No. 1, pp. 4-10.

[5] K. KAIKHAH, "Automatic Text Summarization with Neural Networks", Second IEEE International Conference on Intelligent Systems, JUNE 2004, pp. 40-44

[6] V. Moscato, F. Persia and A. Penta, "iWIN: A Summarizer System Based on a Semantic Analysis of Web Documents", IEEE Sixth International Conference on Semantic Computing, 2012.