Text Summarization:An Overview

Article · October 2013

CITATIONS
14

READS
43,539

3 authors, including:

Samrat Babar
Sanjeevan Engineering and Technology Institute Panhala
3 PUBLICATIONS **110** CITATIONS

SEE PROFILE

Text Summarization:An Overview

Mr.S.A.Babar,M.Tech-CSE,RIT

1.Abstract:

In this new era,where tremondous information is available on the internet,it is most important to provide the improved mechanism to extract the information quickly and most efficiently . It is very difficult for human beings to manually extract the summary of a large documents of text. There are plenty of text material available on the internet. So there is a problem of searching for relevant documents from the number of documents available, and absorbing relevant information from it.In order to solve the above two problems, the automatic text summarization is very much necessary.Text summarization is the process of identifying the most important meaningful information in a document or set of related documents and compressing them into a shorter version preserving its overall meanings.

2.Introduction:

Before going to the Text summarization, first we, have to know that what a summary is. A summary is a text that is produced from one or more texts, that conveys important information in the original text, and it is of a shorter form. The goal of automatic text summarization is presenting the source text into a shorter version with semantics.The most important advantage of using a summary is ,it reduces the reading time. Text Summarization methods can be classified into extractive and abstractive summarization. An extractive summarization method consists of selecting important sentences, paragraphs etc. from the original document and concatenating them into shorter form. An Abstractive summarization is an understanding of the main concepts in a document and then express those concepts in clear natural language.

There are two different groups of text summarization : indicative and informative.Inductive summarization only represent the main idea of the text to the user. The typical length of this type of summarization is 5 to 10 percent of the main text.On the other hand, the informative summarization systems gives concise information of the main text .The length of informative summary is 20 to 30 percent of the main text .

3.Main steps for text summarization:

There are three main steps for summarizing documents.These are topic identification, interpretation and summary generation.

3.1. Topic Identificatio:The most prominent information in the text is identified .There are different techniques for topic identification are used which are Position, Cue Phrases, word frequency.Methods which are based on the position of phrases are the most useful methods for topic identification.

3.2. Interpretation :Abstract summaries need to go through interpretation step. In This step, different subjects are fused in order to form a general content.

3.3. Summary Generation :In this step, the system uses text generation method.

4. **Extractive text summarization :** This process can be divided into two steps: Pre Processing step and Processing step. Pre Processing is structured representation of the original text. It usually includes: a) Sentences boundary identification. In English, sentence boundary is identified with presence of dot at the end of sentence. b) Stop-Word Elimination—Common words with no semantics c) Stemming—The purpose of stemming is to obtain the stem or radix of each word, which emphasize its semantics.

In Processing step, features influencing the relevance of sentences are decided and calculated and then

weights are assigned to these features using weight learning method. Final score of each sentence is determined using Feature-weight equation. Top ranked sentences are selected for final summary.

Summary evaluation is a very important aspect for text summarization. Generally, summaries can be evaluated using intrinsic or extrinsic measures. While intrinsic methods attempt to measure summary quality using human evaluation and extrinsic methods measure the same through a task-based performance measure such the information retrieval-oriented task.

5. TEXT SUMMARIZATION HISTORY:

In the past, extractive summarizers have been mostly based on scoring sentences in the source document. The most common and recent text summarization techniques use either statistical approaches, or linguistic techniques.The high frequency words ,standard keyword ,Cue Method,Title Method ,Location Method are used for weighting the sentences.

6. FEATURES FOR EXTRACTIVE TEXT SUMMARIZATION :

Most of the current automated text summarization systems use extraction method to produce a summary .Sentence extraction techniques are commonly used to produce extraction summaries. One of the methods to obtain suitable sentences is to assign some numerical measure of a sentence for the summary called sentence scoring and then select the best sentences to form document summary based on the compression rate. In the extraction method, compression rate is an important factor used to define the ratio between the length of the summary and the source text. As the compression rate increases, the summary will be larger, and more insignificant content is contained. While the compression rate decreases the summary to be short, more information is lost. In fact, when the compression rate is 5-30%, the quality of summary is acceptable.

7. EXTRACTIVE SUMMARIZATION METHODS :

A. Term Frequency-Inverse Document Frequency (TF-IDF) method:

B. Cluster based method:

C. Graph theoretic approach:

D. Machine Learning approach:

E. Text summarization with neural networks :

F. Automatic text summarization based on fuzzy logic :

7.1 Term Frequency-Inverse Document Frequency (TF-IDF) method:

It is a numerical statistic which reflects how important a word is in a given document.The TF-IDF value increases proportionally to the number of times a word appears in the document.This method mainly works in the weighted term-frequency and inverse sentence frequency paradigm .where sentence-frequency is the number of sentences in the document that contain that term. These sentence vectors are then scored by similarity to the query and the highest scoring sentences are picked to be part of the summary.Summarization is query-specific .

The hypothesis assumed by this approach is that if there are ‘‘more specific words’’ in a given sentence, then the sentence is relatively more important. The target words are usually nouns .This method performs a comparison between the term frequency (tf) in a document -in this case each sentence is treated

as a document and the document frequency (df), which means the number of times that the word occurs along all documents. The TF/IDF score is calculated as follows:

$$TF/IDF(w) = DN\left(\frac{log(1+tf)}{log(df)}\right)$$

where DN is the number of documents.

7.2 Cluster based method:

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

Example:

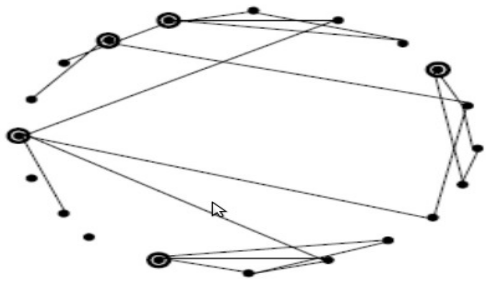
Cluster 1	⟨ambulance, be, scene⟩ ⟨police, be, scene⟩ ⟨ambulance, stream, area⟩ ⟨people, be, street⟩ ⟨pedestrian, peer, sky⟩
Cluster 2	⟨minister, inform, crash⟩ ⟨president, inform, incident⟩ ⟨spokesman, say, minister⟩
Cluster 3	⟨airplane,crash,tower⟩ ⟨plane, strike, building⟩ ⟨clock, fall, floor⟩ ⟨terror, attack, Washington⟩ ⟨terror, attack, New York⟩

Clusters of Triplets

7.3 Graph theoretic approach:

In this technique, there is a node for every sentence . Two sentences are connected with an edge if the two sentences share some common words, in other words, their similarity is above some threshold. This representation gives 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. The second result 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 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

7.4 Machine Learning approach :

In this method,the training dataset is used for reference and 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 from the training data, 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 used in classification. **S** is the summary to be generated, and P (s∈< S | F1, F2, ..., FN) is the probability that sentence s will be chosen to form the summary given that it possesses features F1,F2...FN.

7.4 Text summarization with neural networks:

In this method, each document is converted into a list of sentences. Each sentence is represented as a vector [f1,f2,...,f7], composed of 7 features.

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

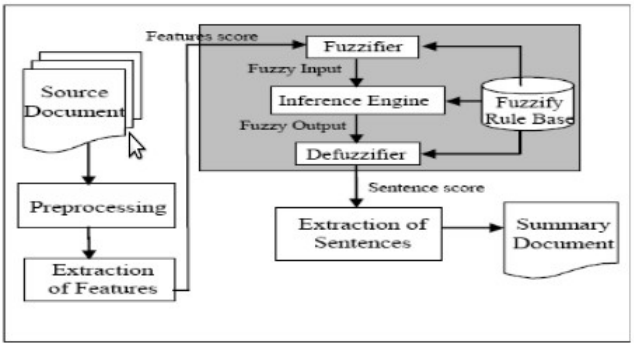
6) f6 Number of thematic words in the sentence

7) f7 Number of title words in the sentence

The first phase of the process involves training the neural networks to learn the types of sentences that should be included in the summary. Once the network has learned the features that must exist in summary sentences, we 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.

7. 5 Automatic text summarization based on fuzzy logic:

This method considers each characteristic of a text such as sentence length, similarity to little, similarity to key word and etc. as the input of fuzzy system .Then, it enters all the rules needed for summarization, in the knowledge base of system. After ward, 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. The obtained value in the output determines the degree of the importance of the sentence in the final summary. The input membership function for each feature is divided into three membership functions which are composed of insignificant values (low L), very low (VL), medium (M), significant values (High h) and very high (VH). The important sentences are extracted using IF-THEN rules according to the feature criteria.



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 IFTHEN 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.

8.Evaluating the summarization systems:

Summary evaluation is a very important aspect for text summarization. Generally, summaries can be evaluated using intrinsic or extrinsic measures. While intrinsic methods attempt to measure summary quality using human evaluation and extrinsic methods measure the same through a task-based performance measure such the information retrieval-oriented task.

Evaluation methods are useful in evaluating the usefulness and trustfulness of the summary. Evaluating the qualities like comprehensibility, coherence, and readability is really difficult. System evaluation might be performed manually(gold standard) by experts .To measure the quality of summary,the manually expert system is used. The qualitative evaluation is done by counting the numbers of sentences selected by the system that match with the human gold standard. To measure the quantittative assessment of the summary the ROUGE evaluator tool is used which consist of precision, recall and F-measure .

9.Conclusion:

Automatic text summarization is an old challenge but the current research direction diverts towards emerging trends in biomedicine, product review, education domains, emails and blogs. This is due to the fact that there is information overload in these areas, especially on the World Wide Web.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 .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.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 into the domain specific areas.