Auto Text Summarization

R. Shelke, Student, Department of Computer Engineering, Marathwada Mitra Mandal’s Institute of technology, Pune  
P. Sarode, Student, Department of Computer Engineering, Marathwada Mitra Mandal’s Institute of technology, Pune  
Ashwini, Student, Department of Computer Engineering, Marathwada Mitra Mandal’s Institute of technology, Pune  
S. Satyan, Student, Department of Computer Engineering, Marathwada Mitra Mandal’s Institute of technology, Pune

ABSTRACT

As the amount of textual Information increases, we come across a need for Automatic Text Summarizers which can provide important textual data from large, well-structured document formats. Automatic summarization is an issue of common concern in computational linguistics and information field, as a computer system of text summarization is an effective means of processing information. In this paper, we proposed various features of Summary Extraction and also analyzed features that are to be applied depending upon the Document.

Keywords

Generic Text Summarizer, Extractive Summary, Information Retrieval.

1. INTRODUCTION

Summary can be defined as a brief and accurate way of representing the important concepts of the given source documents. Humans, during the process of text summarization, understand the concept of source document and create a summary which conveys the essence of the document whereas in automated systems this is a complex task. As the quantity of information available in electronic format continues to grow, research into automatic text summarization has taken huge importance. There are two types of summary Extractive and Abstractive. Abstractive summary represents use of (NLP) whereas Extractive summary is based on copying exact sentences from source document. Presently it is not possible that the computer can understand every aspect behind Natural Language processing. So, our Scope is limited to Extractive based summary.

2. PRESENT THEORY AND PRACTICES

Over the last six decades, the problem of text summarization has been come acrossed from many different perspectives, in different domains. Some of the techniques used for text summarizer are

2.1 Connectionist Approach to Generic Text Summarization

The aim is to auto summarize large documents. This technique utilizes adaptive, incremental learning and knowledge representation system that evolves its structure and functionality. This approach proposes use of Parts of Speech disambiguation using a recurrent neural network, this approach deals with sequential data [2].

2.2 Ranking of Text Units According To Shallow Linguistic Features

This approach finds the most prominent text/sentence using various shallow linguistic features, taking degree of connectedness among the text units into consideration so that it reduces the poor linking sentences in the resulting summary. In this method the effect of lexical chain scoring after the nouns and compound nouns are chained by searching of lexically organized relationships between words in the text format using WordNet and lexicographical relationships such as synonyms and hyponyms For extracting summary all the basis of the sum of the scores of the words in each sentence. The scores of words are decided by using various features such as term frequencies, cue phrases measuring lexical resemblance etc. for scoring the text units using various lexical chains, heuristics and the Vector Space approaches which can decide more correlated sentences in the text[4].

3. PROBLEM STATEMENT

This project describes a system for the summarization of single and multiple documents. The system produces multi as well as single document summaries using clustering techniques for identifying common terms across the set of documents. For each term, the system identifies representative passages that are included in the final summary. Results of our evaluation are also presented.

4. PROJECT SCOPE

There are many aspects and research regarding automatic document summarization that, apart from their importance, cannot be investigated. Our project work is to a wider scope to build a document summarization using Feature extraction methods and Text Ranking Method.

Single-document summarization

In Many scenarios require the summarization of a single document instead of a set of input documents. Single-
document summarization can be viewed as a simplified version of the standard multi-document summarization task, and is therefore addressed by the summarizers developed in this report.

5. SYSTEM FEATURES

This include Functional requirement of system and major services provided by proposed system. Service includes reduction of manual work and time, user friendly interface and cost efficient. For this we have used encryption and decryption for mail and also we can use OTP service for login security. Main priority is customer’s interest, we provide business logic for implementing different aspects for customer. The user will register with username and password. He will be provided the userid which he has registered previously in order to login in our application. When the login process will done user will upload his document for summary.

6. PROPOSED SYSTEM

The aim is to auto summarize documents. This approach utilizes adaptive, incremental learning and that evolves its structure and functionality. This approach proposes use of Parts of Speech disambiguation, and capable of dealing with sequential data.

1. Preprocessing

1.1 Remove stop words

Stop word referred to counter the obvious fact that many of the words contained in a document do not contribute particularly to the description of the documents content they are less frequent and non-relevant words. For instance, words like “the”, “is” and “and” contribute very little to this description. Therefore it is common to remove these stop words in order to the construction of the document summary, leaving only the content bearing words in the text during processing.

1.2 Stemming

Removing suffixes by automatic means is an operation which is especially useful in the field of information retrieval. Terms with a common stem will usually have similar meanings, for example:

CONNECT – CONNECTED – CONNECTING

Frequently, the performance of an IR system will be improved if term groups such as this are conflated into a single term. This may be done by removal of the various suffixes -ED, -ING, -ION to leave the single term CONNECT. The proposed system uses a porter algorithm, and this section shows how the system performs stemming. It is removes about 60 different suffixes, which involves a multi-step process that successively removes short suffixes, rather than removing in a single step the longest suffix.

2. Sentence Ranking and selection

2.1 Statistics: Using simple statistics to determine the summary sentence of a text by using the words weight, weight sequence and sentence frequency.

2.2 Heuristics: Using heuristic features, to determine the summary sentence of a text. Similarity with the title, specify if the word in the sentence occurs in the document’s title or not. Give higher degree to sentence’s that have more words similarity with words in the document’s title.

2.3. Sentence Selection:

The proposed system uses statistics and heuristics method for finding summary, after finding score of each sentence by the specified methods and features the combination function is required to ranking sentences with different weights for giving them the final sentence score. We have used the Sentence Ranking Algorithm which is described below. The original SentenceRank algorithm was described by Lawrence Page and Sergey Brin in several publications. It is given by

PR(A) = (1-d) + d (PR(T1)/C(T1) + ... + PR(Tn)/C(Tn))

where

- PR(A) is SentenceRank of page A,
- PR(Ti) is SentenceRank of pages Ti which link to page A,
- C(Ti) is the number of outbound link on page Ti and
- d is a damping factor it can be set between 0 and 1.

So, first of all, we see that SentenceRank does not rank document as a whole, but is determined for each sentence individually. Further, the SentenceRank of sentence A is
recursively defined by the SentenceRank of those sentences which link to sentence A.

The Sentence Rank of sentence Ti which link to sentence A does not influence the SentenceRank of sentence A uniformly. Within the PageRank algorithm, the PageRank of a sentence T is always weighted by the number of outbound links C(T) on sentence T. This means that the more outbound links a sentence T has, the less will sentence A benefit from a link to it on sentence T.

Finally, the sum of the weighted SentenceRank of all page Ti is then added up. The outcome of this is that an additional inbound link for sentence A will always increase sentence A's PageRank.

The weighted SentenceRank of sentence Ti is then multiplied with a damping factor d which can be set between 0 and 1. Due to that, the scope of SentenceRank benefit for a sentence by another sentence linking to it is reduced.

The characteristics of PageRank can be illustrated by following example.

We regard a small document consisting of three sentence A, B and C, whereby sentence A links to the sentence B and C, sentence B links to sentence C and sentence C links to sentence A. Accordingly the damping factor d is often set to 0.85, but to keep the calculation simple we set damping factor to 0.5. The exact value of the damping factor d confessedly has effects on sentence Rank, but it does not influence the fundamental principles of sentence Rank. So we get the equation as follows:

$$PR(A) = 0.5 + 0.5 \times PR(C)$$
$$PR(B) = 0.5 + 0.5 \times \frac{PR(A)}{2}$$
$$PR(C) = 0.5 + 0.5 \times \left( \frac{PR(A)}{2} + PR(B) \right)$$

These equations can be well solved. We obtain the sentence rank values for the single document as follow:

$$PR(A) = \frac{14}{13} = 1.07692308$$
$$PR(B) = \frac{10}{13} = 0.76923077$$
$$PR(C) = \frac{15}{13} = 1.15384615$$

It is obvious that the sum of all sentences sentence Rank is 3 which is equal to number of sentence in a document. Most of the time, the documents consist of billions of documents and it is not possible to find a solution by inspection.

7. SYSTEM ARCHITECTURE

![Fig. Architecture of System](image)

8. ADVANTAGES

- Works Instantly Reading the entire article, breaking it and separating the important ideas from the original text takes time and effort.
- Improves quality, Some software summarizes not only documents but also web pages.
- This highly improves productiveness as it quicken surfing process.
- Does Not Miss Important Facts.

9. LIMITATIONS

The documents which are in the form of RichText, Pdf and Text format can only auto summarize.

10. EXPECTED RESULT

The Hopefull result of this project is that the client will get an application that will execute on client side and get the summary of the input document as per his/her requirement.

REFERENCES
