Text Summarization Basing on Font and Cue-Phrase Feature for a Single Document

S.V.S.S. Lakshmi[1], K.S. Deepthi[1], and Ch. Suresh[2]

[1] Department of Computer Science and Engineering, ANITS,
Sangivalasa, Visakhapatnam-531162, AP, India
lakshmi.it@anits.edu.in
[2] Department of Information Technology, ANITS,
Sangivalasa, Visakhapatnam-531162, AP, India

Abstract. In recent times owing to the magnitude of data present digitally across networks over a wide range of databases the need for text summarization has never been higher. The following paper deals with summarization of text derived from the syntactic and semantic features of the words in the document. We apply the technique of calculating a threshold value from both the attribute and semantic structure of the individual words. The algorithm helps in calculating the threshold value in order to give weightage to a particular word in a document. Initially the document undergoes the preprocessing techniques; the obtained data will be kept in a data set, then on that data we will apply the proposed algorithm in order to get a summarized data.

Keywords: summarization, threshold, semantic, extraction, cue word, statistical, preprocessing.

1 Introduction

In recent times the documents we see and come across are exceeding our capacity and time frame to read them. Hence the need for automatic text summarization has never been greater. This enables us to understand and comprehend the idea and thought behind a document in a compressed form. For this purpose we extract sentences and form an effective summary using our proposed algorithm. The summary that is obtained remains unchanged in its meaning. The summary should meet the major concepts of the original document set, should be redundant-less and ordered. The summarization of text is lies on sentences either being extracted or generated. Extracting a sentence basically derives upon accessing words that have peculiar attributes i.e. italic, bold, and underlined, title word, start word etc.

Generating sentences has possessed far too many problems owing to its complex analytical nature and profound knowledge required for calculating and generating words basing on the documents provided. The paper focuses on extracting sentences based upon the syntactic and semantic features of the content in the text. It is argued that summaries generated automatically are far more inferior to the ones generated by humans.

S.C. Satapathy et al. (eds.), *Emerging ICT for Bridging the Future – Volume 2,*
Advances in Intelligent Systems and Computing 338, DOI: 10.1007/978-3-319-13731-5_58

1.1 Summarization

Text mining is the analysis of data contained in natural language text. Text mining works by transposing words and phrases in unstructured data into numerical values which can then be linked with structured data in a database and analyzed with traditional data mining techniques.Text summarization (TS) is the process of identifying the most salient information in a document or set of related documents and conveying it in less space (typically by a factor of five to ten) than the original text. Identifying the redundancy is a challenge that hasn't been fully resolved yet. Method consists of selecting important sentences, paragraphs etc. from the original document and concatenating them into shorter form without changing the original concept or meaning of the document.It is very difficult to achieve consistent judgments about summary quality from human judges. This fact has made it difficult to evaluate (and hence, improve) automatic summarization.

Summarization task is done in two different methods, i.e. extractive and abstractive. An extractive summarization of sentences is decided based on statistical and linguistic features of sentences. An Abstractive summarization [9] [10] attempts to develop an understanding of the main concepts in a document and then express those concepts in clear natural language. It uses linguistic methods to examine and interpret the text and then to find the new concepts and expressions to best describe it by generating a new shorter text that conveys the most important information from the original text document.

1.2 Summarization Features

The **font based feature** i.e. bold, italic, underlined and all the combination of these are considered to be more important when calculating the weight for ranking the sentences of the document. For this reason the accuracy rate of our system is more than that of MS-Word automatic text summarization in most cases. **cue-phrase feature i**s based on the hypothesis that the relevance of a sentence is computed by the presence or absence of certain cue words in the cue dictionary. Sentences containing any cue phrase (e.g. “in conclusion”, “this letter”, “this report”, “summary”, “argue”, “purpose”, “develop”, “attempt” etc.) are m**ost** likely to be in summaries**.Sentence location feature** is usually first and last sentence of first and last paragraph of a text document are more important and are having greater chances to be included in summary.**Sentence length feature** is **v**ery large and very short sentences are usually not included in summary**.Proper noun feature** is name of a person, place and concept etc. Sentences containing proper nouns are having greater chan**ces** for including in summary**.Upper-case word feature** finds **s**entences containing acronyms **o**r proper names are included**.**Title method finds the sentence weight is computed as a sum of all the content words appearing in the title and (sub-) headings of a text. you have more than one surname, please make sure that the Volume Editor knows how you are to be listed in the author index.

Single document summarization is the process of creating a summary from a single text document. Multi- document summarization shortens a collection of related documents; into single summary.The proposed algorithm mainly focuses on single document summarization. If we apply each algorithm separately i.e. font based feature and cue-phrase feature we will get summarized data with less accuracy, so we propose an algorithm which is combination of both the above mentioned algorithms in order to get summarized document with maximum accuracy.

2 Proposed System

The above mentioned techniques can individually produce summarized data but with less accuracy. Different combinations of algorithms were developed to produce summarized data but they could not able to produce with maximum accuracy.so we have developed a methodology which is combination of font feature based and cue-phrase based feature, So our algorithm succeed to produce summarized data with maximum accuracy.The steps are following :

a) Select the Features (FONT & CUE WORD)

b) Identify the Tokens based on Threshold value & assign the ranks

c) Select the Sentences Based on ranking

d) Generate a Summary

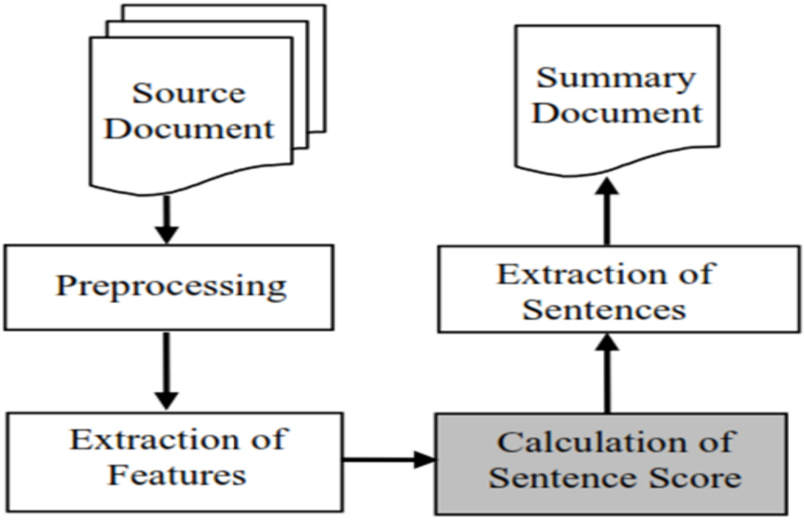

Fig. 1. Text summarization based on Features

The summarized data will be obtained by using an algorithm which is combination of both font based and cue-phrase technique. Initially the document undergoes preprocessing techniques then we apply proposed technique on the obtained data after preprocessing.

2.1 Threshold Value Calculation

A value beyond which there is a change in the manner is called Threshold value. Where T_Value is taken as follows:

- For F b, F I, F u, F c, **T_Value=1**
- For F b – F I , F I – F u , F b – F u , F b –F c , F I -F c ,F u - Fc **T_Value=2**
- For F b –F u –F c, F b – F I – F c , F I - F u – F c **T_Value=3**
- For F b – F I - Fu- F c **T_Value =4**

2.2 Algorithm

Input: A text in .txt or .rtf format.

Output: A relevant summarized text which is shorter than the original text

Step 1: Read a text in .txt or .rtf format and split it into individual tokens.

Step 2: Remove the stop words to filter the text.

Step 3: Add a weight to the sentences which are appear in bold, italic, underlined, cue word or any combination of these. The weight value can be calculated as:

F = (∑ (Frequency of the Special Term * T_Value)) / Total No. of Special terms in the sentence

Step 4 : Rank the individual sentences according to their Weights. If the F value is high then the rank is lower.

Step 5: Finally, extract the higher ranked sentences including the first sentence of the first paragraph of the input text in order to find the required summary.

3 Results and Discussion

We had tested our algorithm by passing 20 documents; we got summarized data with different accuracy. The average accuracy we got is 75%. The graph is shown below:

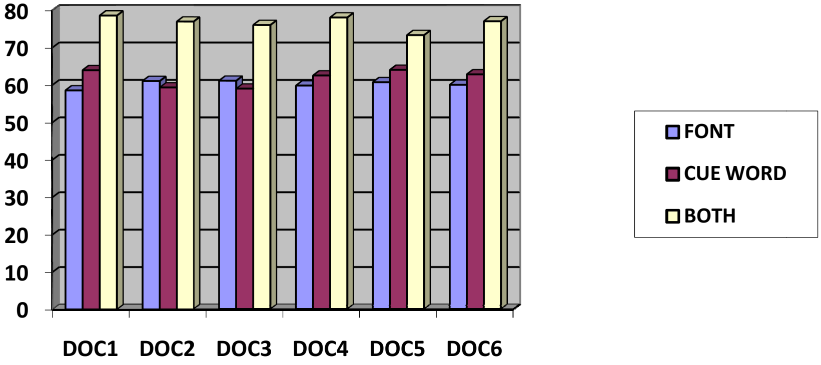

Fig. 2. Summarization Accuracy Graph

4 Conclusion and Future Work

We had presented an algorithm which produces a summarized data using combination of font based and cue- phrase algorithm. The algorithm produces an efficient summarized data when compared to word pruning and using individual algorithms of above mentioned. Finally we could produce accurate summarized data using our proposed methodology.

As our proposed algorithm is combination of above mentioned two techniques, so we could produce able to produce summarized data with 75% accuracy.so future work is to generate summary using abstractive method and we can also use all the features which are mentioned above to generate a summary using extractive approach to get 100% accurate summarized data .

References

1. Agarwal, R.: Semantic features extraction from technical texts with limited human intervention. Ph.D Thesis, Mississippi State Univ, U.S. (1995)
2. D' Avanzo, E., Magnini, B., Vallin, A.: ITC-irst. Keyphrase Extraction for Summarization purposes: The LAKE system at DUC-2004. In: Document Understanding Conference (2004)
3. Edmundson, H.P.: New methods in automatic extracting. Journal of the Association for Computing Machinery 16(2), 264–285 (1969)

4. Kruengkrai, C., Jaruskulchai, C.: Generic Text Summarization Using Local and Global Properties. In: Proceedings of the IEEE/WIC International Conference on Web Intelligence (2003)
5. Jezek, K., Steinberger, J.: Automatic Text Summarization (the state of the art 2007 and new challenges). In: Znalosti 2008, pp. 1–12 (2008)
6. Salton, G., Buckley, C.: Term-weighting approaches in automatic text retrieval. Information Processing and Management 24, 513–523 (1988); Reprinted in: Sparck-Jones, K., Willet, P. (eds.) Readings in I. Retrieval, pp. 323–328. Morgan Kaufmann (1997)
7. Aliguliyev, R.M.: A new sentence similarity measure and sentence based extractive technique for automatic text summarization. Expert Systems with Applications 36(4), 7764–7772 (2009)
8. Fattah, M.A., Ren, F.: Automatic Text Summarization. Proceedings of World Academy of Science, Engineering and Technology 27, 192–195 (2008) ISSN 1307- 6884
9. Hariharan, S.: Multi Document Summarization by Combinational Approach. International Journal of Computational Cognition 8(4), 68–74 (2010)
10. Wasson, M.: Using Leading Text for News Summaries: Evaluation results and implications for commercial summarization applications. In: Proceedings of the 17th International Conference on Computational Linguistics and 36th Annual Meeting of the ACL, pp. 1364–1368 (1998)