IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661, p- ISSN: 2278-8727Volume 9, Issue 3 (Mar. - Apr. 2013), PP 62 -66
www.iosrjournals.org

An automatic Text Summarization using feature terms for relevance measure

Anita.R.Kulkarni[1], Dr Mrs. S.S.Apte[2]
[1](Computer Science & Engg Department, Walchand Institute of Technology, India)
[2](Computer Science & Engg Department, Walchand Institute of Technology, India)

Abstract: *Text Summarization is the process of generating a short summary for the document that contains the overall meaning. This paper explains the extractive technique of summarization which consists of selecting important sentences from the document and concatenating them into a short summary. This work presents a method for identifying some feature terms of sentences and calculates their ranks. The relevance measure of sentences is determined based on their ranks. It then uses a combination of Statistical and Linguistic methods to identify semantically important sentences for summary creations. Performance evaluation is done by comparing their summarization outputs with manual summaries generated by three independent human evaluators.*

Keywords: - Generic text summarization, Relevance measure, Semantic Analysis, Term-frequency Rank.

I. INTRODUCTION

With enormous growth of information on WWW, Conventional IR techniques have become inefficient for finding relevant information effectively. Given a keyword-based search on the internet, it returns thousands of documents overwhelming the user. It becomes very difficult and time consuming task to find the relevant documents.. Therefore this paper provides text summarization approach as a solution to this problem. This approach reduces the time required to find the web document having relevant and useful data. Text Summarization is the process of automatically creating a compressed version of the given text. This compressed version is called summary. Text summarization has two approaches, namely Extraction and Abstraction. This paper focuses on extractive summarization.

Text summaries can be either query relevant or generic summaries. Query relevant summaries contain sentences or passages from the document that are query specific. It is achieved by using conventional IR techniques. On the other hand, generic summary provides an overall sense of the document's content. In this method neither query nor any topic will be provided to summarizer. It is a big challenge for a summarizer to produce a good quality generic summary. In this paper, we propose an extractive technique for text summarization by using feature terms for calculating the relevance measure of sentences and extract the sentences of highest ranks. Then we perform their semantic analysis to identify semantically important sentences for creating a generic summary. Our proposed work generates a generic summary .There are various techniques that have been applied in text summarization. It includes

1. Statistical approach
2. knowledge-based approach
3. Linguistic Technique

I.1 Statistical approach

The statistical approach summarizes using statistical features. It uses Information Retrieval methods to determine the relevance of sentences. IR methods use frequency of terms and phrases to determine the relevance of the sentence. However IR-based methods do not give satisfactory results as the summary generated is not very coherent. Recently classification and position-based methods are also used for summarization. The classifier uses the training data to classify the sentences as relevant or not. An algorithm using certain parameters is used for this task. The position-based methods use the position of sentences to determine whether the given sentence is important to be included in the summary or not. Statistical approaches are faster as they are domain-independent but do not produce good quality of summary.

I.2 knowledge-based Approach

The knowledge-based approach stands in direct contrast to the statistical approach. This approach interprets the text using extensive domain knowledge as well as natural language techniques and then summarizes it.

I.3 Linguistic approach

The natural language processing techniques are the other ways to produce an abstracted summary by understanding the context of the original content. The techniques of natural language processing & statistical approach can be combined to generate more useful and meaningful summaries.

This paper focuses on method that utilizes both types of text summarization. In our study we focus on sentence based extractive summarization. This technique attempts to identify a set of sentences that are most important for the overall understanding of the given document. This paper presents an approach to generic summarization on single document using both statistical method and linguistic method. The rest of the paper is organized as follows. Section 2 gives the related work, section 3 gives the explanation of the method, section 4 gives Evaluation methods section 5 gives conclusion and section 6 gives future scope

II. Related work

The earliest instances of research on summarizing scientific documents proposed paradigms for extracting salient sentences from text using features like word and phrase frequency, position in the text and key terms or key phrases [1].

Most of the research work has focused on extraction in late 80s.It focused on extracts rather than abstracts along with the renewed interest in earlier surface level approaches. Evolution of different features of sentences and their extraction for sentence scoring has been studied in various research papers [1][2].

Other significant approaches such as hidden Markov models and log-linear models to improve extractive summarization were studied [3][4].

Various works Published has concentrated on different domains where text summarization is used. The domain-specific text summarization then became popular which used corpus for keyword frequencies [4][7]. This emphasizes on extractive approaches to summarization using statistical methods.

Recent Papers published have shown the use of fuzzy logic and neural networks [5] for text summarization in order to improve the quality of the summary created by the general statistical method, they proposed fuzzy logic based text summarization. They have also proposed an improved feature scoring technique based on fuzzy logic for producing good summary [8]. They have proposed to address the problem of inaccurate and unsure feature score utilizing fuzzy logic.

A neural network was used for summarizing news articles in the recent work[9]. A neural network was trained to learn the significant features of sentences that are suitable for inclusion in the article summary. Then the significant features are generalized and combined and modified accordingly. Then the neural network acts as a filter and summarizes news articles.

We also discuss about some summarization tools

1 SweSum[1] a summarization tool from Royal Institute of Technology, Sweden.

2 MEAD- a public domain multi-lingual multi-document summarization system developed by the research group of Dragomir Radev

3 LEMUR[2]- a summarizer toolkit that provides summary with its own search engine.

2.1 SWESUM

It is an online summarizer [1] that was first constructed by Hercules Dalianis and developed by Martin Hassel. It is a traditional extraction-based domain specific text summarizer that works on sentences from news text using HTML tags. For topic identification, Swesum makes use of hypothesis, where the high frequent content words are keys to the topic of the text. Sentences that contain keywords are scored high [5]. Sentences that contain numerical data are also considered to carry important information. These parameters are put into a combination function with modifiable weights to obtain total score of each sentence. It is completely user dependent and it is also difficult for inexpert user to set the parameter of the SweSum.

2.2 MEAD

The centroid-based method [6] is the most popular extractive summarization method. MEAD is an implementation of this method. MEAD uses three features to determine the rank of the sentence. They are centroid-score, position and overlap with first sentence. It computes centroid using tf-idf-type data. It ranks candidate summary sentences by combining sentence scores against centroid, text position value, and tf-idf-title. Sentence selection into summary is constrained by summary length and redundant new sentences avoided by checking cosine similarity against prior ones. It only works with news text, but not with web pages whose structure is different from news articles

2.3 LEMUR

It is a toolkit [2] used for searching the web and it makes summary of single document and multidocuments [7]. It uses TF-IDF (vector model), Okapi (Probabilistic model) for multidocument summarization and standard query language as relevance feedback. Lemur also provides standard tokenizer that has options for stemming and stop words.

III. A better approach to summarization

Our work uses a combination of statistical and Linguistic [4] method to improve the quality of summary. It works in four phases

i) Preprocessing of text
ii) Feature extraction of both words and sentences
iii) Summarization algorithm for calculation of rank using features score.
iv) Extracting sentences of higher ranks to generate summary

The figure 1 below shows the architecture of this technique.

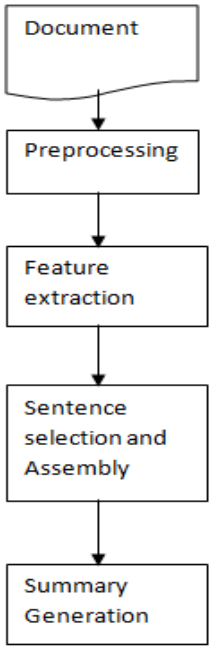

Figure 1

III.1 PREPROCESSING OF TEXT:

It involves 4 steps

- Sentence Segmentation
- Tokenization and POS tagging
- Removing Stop Word
- Word Stemming

The system accepts the document from DUC 2002 and divides it into sentences using sentence segmentation.

Then it is fed to the standard parser for generating tokens. A parser cum POS tagger provided by Stanford is used to tag the input text into various parts of speech such as nouns(NN), verbs(VBZ), adjectives(JJ) and adverbs(ADVB), determiners(DT) coordinating conjunction(CC) etc. It also divides the text into groups of syntactically correlated parts of words as Noun phrase[NP], verb phrase[VB], adjective phrase[AP] etc

Example: The rose has a variety of colors, shapes and sizes.

[NP the/DT rose/NP] [VP has/VBZ] [NP a/DT variety/NN [PP of/of] [NP colors/NN shapes/NN and/CC sizes/NN]

Next, Stop Words are removed. Stop words are the words which appear frequently in document but provide less meaning in identifying the important content of the document such as ‘a’, ‘an’, ‘the’, etc..

The last step for preprocessing is Word Stemming; Word stemming is the process of removing prefixes and suffixes of each word.

III.2 FEATURE EXTRACTION

The text document D, after preprocessing is subjected to feature extraction by which each sentence in the text document obtains a feature score based on its importance. Each feature is given a value between 0 and 1. The important text features to be used in this d system are as follows:

III.2.1 Title feature

The sentences that contain words or their synonyms, in the title should be considered for inclusion in the summary as they reveal the theme of the document**.**

III.2.2 Term frequency-Inverse Sentence Frequency

Term frequency TF (t, d) of term t in the document d is defined as the number of times that term t occurs in d.

Inverse Sentence frequency is used to measure the information content of a word. It says that terms that occur in most of the sentences are less important than the ones that occur in few sentences.

Inverse sentence frequency is calculated as follows

$$ISF_i=\log (N/ni)$$

Where N denotes the number of sentences in the document and ni denotes the number of sentences in which term i occurs.

$$TF\text{-}ISF_i=TF*\log (N/ni)$$

The TF-ISF of all terms in a sentence is added .The sentence having TF-ISF greater than a threshold is selected for inclusion in summary.

III.2.3 Existence of Indicated words

Indicated words are the information containing words that help to extract important sentences. They can be domain-specific words. For example, if the domain of text summarization is research articles or papers, then following are some of the indicated words.

Purpose: It gives the information of need of research work or the motivation for research work.

Methods: It indicates the method or experimental procedures used in research

.

Conclusions: This word indicates the significance of research

The sentences containing indicated words are considered important to be included in summary.

The list of indicated words is predefined.

III.2.4 Existence of Cue-Phrase

Sentences containing cue phrase such as “This letter” “this paper” ,”The proposed work”, “this report” ,”develop” etc are candidate sentences to be included in the summary.

III.2.5 Sentence Position

The first sentence and the last sentence of the paragraphs are usually included in the summary.

III.2.6 Existence of Key phrase

The sentences that contain noun phrase or verb phrase are considered important to be included in summary. This is possible by noun and verb chunking. The n-best output for taggers could be used to define chunks. This makes POS Tagging at Lexical level.

III.2.7 Correlation among sentences

Correlation of sentences is very important for the summary as the sentence often refers to the previous or the next sentence. If we consider only the relation of a sentence with the previous sentence then sentences starting with connectives such as such, although, however, moreover ,also, this, those and that are related with the preceding sentence. In such case the preceding sentence is also selected to be included in the summary. If the rank of the preceding sentence is equal to or greater than 70% of the rank of the selected sentence, then it is included in the summary.

IV. Sentence Selection And Assembly

The score of every feature will be normalized between 0 and 1 and the score of the sentence is the sum of all the scores of every feature. The score of the sentence is called the rank of the sentence. The sentences are

stored in descending order of their ranks and top n highest scoring sentences are considered for summary, where value of n is based on choice for percentage of summary.

V Summary Generation

The sentences are put into the summary in the order of their positions in the original document. URLs and E-mails are removed from them as they do not contain important information.

VI. Evaluation

Evaluation is a key part of any research and development effort. Each approach should have an evaluation. It will not only tell how effective of the approach, but also can be used to study and improve sentence selection criteria.

Text summarization can be evaluated by using precision and recall, which are well known measurable quantities based on statistical approach in the information retrieval discipline. Precision refers to the measure of correctness of output based on relevance of the retrieved information. Recall measures the completeness of the output, which refers to the relevant extracted information. Relevant sentences are those that occur in summary generated by human experts and retrieved sentences are those that are selected by the summarization system.
The harmonic mean of precision and recall is called as F-measure. All these 3 parameters will be calculated for generated summaries of test documents.

VII. Conclusions

In this paper we explained an approach to summarize a single document using statistical and Linguistic approaches. We calculated scores of word and sentence features. Then we calculated the rank of the sentences by summing up these scores. The top n ranked sentences were picked up to be included in summary. Some minor post processing was done on these sentences to generate the final summary.

VIII. Future Scope

This work focuses on single document summarization. It can be extended to multidocument and multilingual summarization

References

[1] H. Dalianis, "SweSum-A Text Summarizer for Swedish", Technical Report, TRITA-NA-P0015, IPLab-174, KTH NADA, Sweden, 2000.
[2] N. McCracken, "IR Experiments with Lemur", Available at: http://www-2.cs.cmu.edu/~lemur/ [December 16, 2009]
[3] T. Chang and W. Hsiao, "A hybrid approach to automatic text summarization", 8th IEEE International Conference on Computer and Information Technology (CIT 2008), Sydney, Australia, 2008.
[4] Ghadeer Natshah, Yasmeen Ta'amra, Bara Amar and Manal Tamini, "Text Summarization: Using combinational Statistical and Linguistic Methods"
[5] Vishal Gupta and Gurpreet Singh Lehal "A survey of Text summarization techniques "Journal of Emerging Technologies in Web Intelligence VOL 2 NO 3 August 2010.
[6] Suneetha Manne and S.Sammen Fatima's"A feature Terms based Method for Improving Text summarization with supervised POS Tagging".
[7] R.Shams, A.Elsayed and Q.M Akter, "A corpus-based evaluation of a domain-specific text to knowledge mapping prototype", A special issue of Journal of Computers, Academy Publisher, 2010(In Press)
[8] T. Chang and W. Hsiao, "A hybrid approach to automatic text summarization", 8th IEEE International Conference on Computer and Information Technology (CIT 2008), Sydney, Australia, 2008.
[9] Stanford NLP Group, "Stanford log-linear part of speech tagger", Available at: Http://nlp.stanford.edu/software/tagger.shtml [June 15, 2009]
[10] Chengcheng L.(2010).Automatic Text Summarization Based On Rhetorical Structure Theory. IEEE. 2010 International Conference on Computer Application and System Modeling (ICCASM 2010).