A New Method of Text Summarization

Anjana Polgampala,
Sri Lanka Institute of Information Technology,
New Kandy Road, Malabe,
Sri Lanka.
+92712079865
appolgampala1992@gmail.com

Abstract-**Today, the tremendous information is available on the internet, is difficult to get the information fast and most efficiently. Text summarization technique deals with the compression of large document into shorter version. it is a process that reduces the size of the text document and extract significant sentences from a text document. This is a novel technique for text summarization. The technique based on local and global properties of words and identifying significant words. The local property of word can be considered the sum of normalized term frequency multiplied by normalized number of sentences containing that word. Global property can be thought of as maximum semantic similarity between a word and title words. Extraction based text summarization involves selecting sentences of high rank from document based on word and sentence features and put them together to generate summary.**

1. INTRODUCTION

As the amount of information grows rapidly, text summarization is getting more important. Text Summarization is a tool to save time and to decide about reading a document or not. It is a very complicated task. It should manipulate a huge quantity of words and produce a cohesive Summary. The main goal in text summarization is extracting the most important concept of text document. Two kinds of text summarization are: Extractive and Abstractive. Extractive method selects a subset of sentences that contain the main concept of text. In contrast, abstractive method derives main concept of text and builds the summarization based on Natural Language Processing. Our technique is based on extractive method. There are several techniques used for extractive method. Some researchers applied statistical criterions. Some of these criterions include TF/IDF (Term Frequency-Inverse Document Frequency), number of words occurring in title , and number of numerical data . Using these criterions does not produce a reader friendly summary. As a result NLP (Natural Language Processing) and lexical cohesion are used to guarantee the cohesion of the summary. In this document, we present a new technique which benefits of the advantages of both statistical and NLP techniques and reduce the number of words for Natural Language Processing. We use two statistical features: term frequency normalized by number of text words and number of sentences containing the word normalized by total number of text sentences.

2. PROPOSED TECHNIQUE – SUMMIT

The goal in extractive text summarization is selecting the most relevant sentences of the text. One of the most important phases in the text summarization process is identifying significant words of the text. Significant words play an important role in specifying the best sentences for summary. I use the advantages of several techniques to make text summarization system better. My technique has five main steps: preprocessing, calculate words score, calculate sentence score and sentence selection.
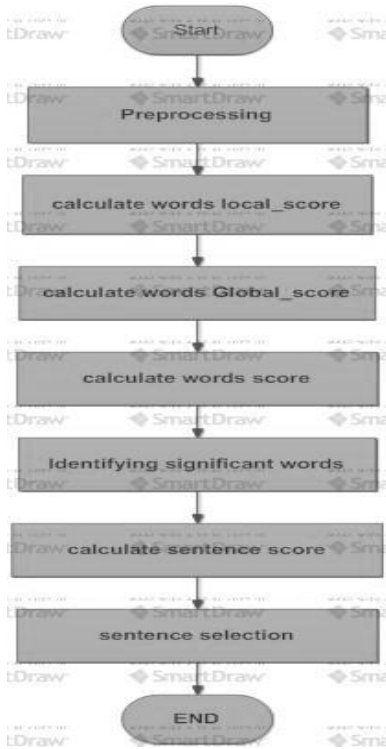
All these steps are shown in Fig 1.

Fig 1: The flowchart of proposed technique

2.1. *Preprocessing*

First of all we perform sentence segmentation to separate text document into sentences. Then sentence tokenization is applied to separate the input text into individual words. Some words in text document do not play any role in selecting relevant sentences of text for summary. Such as stop words. Finally, we separate nouns of the text document. My technique works on nouns of text.

2.2. *Calculating Local score of words*

In this step we utilize combination of statically criterions and lexical cohesion to calculate text words scores. Finding semantic relation between words is a complicated and time consuming process.so first of all we remove unimportant words.

In this phase, we use two criterions. Term frequency of the word(represented by TF) and number of sentences containing the word(represent by sen_count).We combine these two criterions to define equation(1) to calculate local score of words.

$$Word\_local\_score = TF * sen\_count \quad (1)$$

2.3. *Calculating global score of words*

In this phase, we consider semantic similarity between text words and title words. We use Wordnet,a lexical database to determine semantic relations between text words and title words.

We define equation (2) to calculate global score of words.

$$Word\_global\_score = Max\ (sim(w,t)) \quad (2)$$

According to this equation. First of all, we calculate the maximum similarity between each word and title words. Then the sum of maximum similarities is calculated to determine global score of words. This score is used in next section

2.4. *Calculating word score*

The final phase in this step is calculating word score.In our technique.word score is calculated by combination of local score and global score of word.we define equation (3) to calculate word score.

$$Word\_score=word\_local\_score + word\_global\_score \quad (3)$$

2.5. *Calculating sentence score*

In this step.we use significant words determined by previous step to calculate sentence score.my technique in this phase is based on Kruengkrai and Jarukuluchi approach, but I changed the parameters. They combined local and global properties of a sentence to determine sentence score as follow:

$$Sentence\_score =\ G + L \quad (4)$$

Where G is the normalized global connectivity score and L is the normalized local score.it results this score in the range of (0,1).

$$G = \sum sim(w,t) \text{ / Total number of words in the sentence} \qquad (5)$$

$$L = \text{number of significant words of a sentence /Total number of significant words of text} \qquad (6)$$

Where sim(w,t) is the maximum semantic relation among sentence words and title and keywords. We consider semantic relation among sentence words and title words to determine the global property of a sentence.

2.6. Sentence selection

After calculating scores of the sentences, we can use these scores to select the most important sentences of text. This is done by ranking sentences according to their scores in decreasing order. Sentences with higher score trend to be included in summary more than other sentences of the text document. In my technique these sentences have more similarity to title. This similarity is measured according to statistical and semantic techniques used in my technique.

3. RESULTS AND DISCUSSION

The proposed system is a novel method of text summarization. The originality of technique lies on exploiting local and global properties of words and identifying significant words. We can get accurate summary within few seconds.

4. CONCLUTION and FUTURE WORK

In this article I proposed a new technique to summarize text documents and research papers. I introduced a new approach to calculate words scores and also identifying significant words of the text.

In future work, I intend to use other features, such as to change the words local score and global score according to readers needs. And also I hope to improve this technique as Abstractive based summarization. As our proposed system summarized only the English texts, this can be improved to summarize the Sinhala and Tamil texts also.

REFERENCES

[1] C.Y.Lin, "Training a Selection Function for Extraction", In Proceedings of eighth international conference on Information and knowledge management, Kansas City, Missouri, United States, pp.55-62,1999.

[2] M.Hoey, Patterns of Lexis in Text. Oxford: Oxford University Press, 1991

[3] K.Svore, L.Vanderwende, and C.Bures, "Enhancing Single-document Summarization by Combining Ranknet and Third-party Sources", In Proceeding of the EMNLPCoNLL.

[4] F.Kyoomarsi, H.Khosravi, E.Eslami, and P.Khosravyan Dehkordy, "Optimizing Text Summarization Based on Fuzzy Logic", In Proceedings of Seventh IEEE/ACIS International Conference on Computer and Information Science, IEEE, University of shahid Bahonar Kerman,2008,pp.347-352.

[5] V.Qazvinian, L.Sharif Hassanabadi, R.Halavati, "Summarization Text with a Genetic Algorithm-Based Sentence Extraction", International of Knowledge Management Studies (IJKMS),2008,vol.4,no.2,pp.426-444.

[6] S.Hariharan, "Multi Document Summarization by Combinational Approach", International Journal of Computational Cognition, 2010, vol.8, no.4, pp.68-74.

[7] M.Abdel Fattah, and F.Ren, "Automatic Text Summarization", Proceedings of World of Science, Engineering and Technology,2008,vol.27,pp.195-192.

[8] L.Suanmali, M.Salem, N.Binwahlan and Salim, "sentence Features Fusion for Text Summarization Using Fuzzy Logic", IEEE, 2009,pp.142-145.