Text Summarization Based on Several Natural Language Techniques

Abeer Khalid Ahmad
Depertment of Computer Science, College of Science, Al-Nahrain University/Baghdad
Email: aabeeeeraa@yahoo.com

Received on: 3/3/2013 & Accepted on: 11/6/2013

ABSTRACT

Because of the great amount of information that provided by internet technologies, the automatic text summarization have become more important. This paper describes a method for summarizing English text. It depends on extractive summarization. The method implies many techniques of statistics and linguistic approaches especially based on morphological rules. The linguistic approaches in this method also include synonym, word-frequencies, word position, and part of speech. It will be shown that merging many statistics and linguistic approaches in one system, gives high accurate results at low threshold values. The system is tested to find the best threshold value, and it was 60%.

Keywords: Keywords Extraction, Morphological Rules, Extractive Summarization, Abstractive Summarization.

تلخيص النصوص بالاعتماد على عدة تقنيات لغوية

الخلاصة

نظرا للكمية الكبيرة من المعلومات التي تقدمها تقنيات الإنترنت، أصبح تلخيص النص بصورة تلقائية أكثر أهمية.هذا البحث يصف طريقة لتلخيص النص الانكليزي. تعتمد الطريقة على التلخيص الاستخراجي. تنطوي الطريقة على العديد من الاسالب الاحصائية واللغوية وخاصة تلك التي تعتمد على قواعد تشريح الكلمات. الاساليب اللغوية في هذه الطريقة تشمل أيضا المرادف، تكرار الكلمات، موقع الكلمة، وجزء من الكلام. سيظهر أن دمج العديد من الاسالب الإحصائية والنهج اللغوية في نظام واحد، يعطي نتائج ذات دقة عالية في قيم عتبة منخفضة. تم اختبار النظام للعثور على أفضل قيمة عتبة، وكانت 60٪.

INTRODUCTION

The subfield of summarization has been investigated by the Natural Language Processing (NLP) community for nearly the last half century. The most important advantage of using a summary is its reduced reading time [1]. Summary is defined as "a text that is produced from one or more texts, it conveys important information in the original text(s), and it is no longer than half of the original text(s) and usually significantly less than that" [2, 3]. A good summary system should reflect the diverse topics of the document while keeping redundancy to a minimum [1].

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. The importance of sentences is decided based on statistical and linguistic features of sentences. An abstractive summarization 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 [2, 1].

The proposed system uses several natural language techniques to gain the aim of text summarization. This system is an extractive English text summarization. Most of the techniques in summarization use simple keyword/phrase extraction or extractive sentence selection methods [4, 5]. A previous work performs the task of keywords extraction, based on several techniques especially, English Morphological Analyzer (EMA) [6]. EMA technique implies many linguistic approaches such as; synonym, word-frequencies, word position, and part of speech. Implying many linguistic approaches gives EMA technique more accurate results [6]. The proposed system collects important sentences from original text. Any statement contains at least one keyword, is considered as important sentence. This system selects important sentences according to the EMA technique (use EMA technique to extract keywords). A heuristic value is assigned for each important sentence, and then summarization will be generated from important sentences with the highest heuristic.

RELATED WORKS

Dr. Ahmed Tariq Sadiq and Noor Amjed Hassan (2008) [7], they proposed several stages to generate summary. Three main disciplines that are integrated to give accurate results: Statistical methods, linguistically approaches using natural language processing and machine learning using association rule. Their system used for integrating text summarization and data mining. The output of their system is an extract from the original text and it saved the meaning of the original text. The sentences of output text are (25-40) % from the sentences of the original text. From the proposed system experiments a good results obtained about 95% for text summarization compared with text summarized get manually by human.

Ahmed Tariq Sadiq, Saran Akram Chawishly and Kanar Shukr Muhamad (2009) [8], they described hybrid system for automatic text summarization which combines statistic

and heuristic methods. As with the statistic method the summary is found based on some statistics features, and with the statistic and heuristic the summary is found based on combined statistic features and heuristic features like word frequency, position, length of sentences, and similarity with the document title. The core of their system is to find the best parameter values of each used features for all used summarization rates.

Ahmed Tariq Sadiq and Enas Tariq Khuder (2010) [9], they presented a good multiple techniques for English text summarization. They use statistical, heuristics, linguistics and machine learning techniques to summarize the text. Statistically, they use the classical measures in the text summarization such as words frequency, cue frequency. Heuristically, they use the title's words, position of words…etc. Linguistically, they use the natural languages processing tools such as part-of-speech, Noun-Phrase-chunk and n-grams. As a machine learning technique they use association rules extraction to find the relational words in different documents. These four techniques executed on 20 different documents to summarize (20-40)% of original document, they have 96% as an acceptable ratio compared with reference human summary.

Naresh Kumar Nagwani and Shrish Verma (2011) [10], they designed and implemented frequent term based text summarization algorithm. The designed algorithm works in three steps. In the first step, the document which is required to be summarized is processed by eliminating the stop word and by applying the stemmers. In the second step, term-frequent data is calculated from the document and frequent terms are selected, for these selected words the semantic equivalent terms are also generated. Finally in the third step, all the sentences in the document, which are containing the frequent and semantic equivalent terms, are filtered for summarization.

THE PROPOSED SYSTEM

The proposed system implies many techniques of statistics and linguistic approaches especially based on morphological rules. This system collects the important sentences to construct the summarized text. The proposed system uses natural language techniques. Firstly, the proposed system extracts all keywords from original text. If a sentence has at least one keyword, then it will be important sentence. The proposed system consists of four main processes; keyword extraction, Sentence Extraction, Sentence Heuristic, and Summarized Text Generation (as illustrated in figure 1). This system uses the EMA technique to gain the keywords. EMA technique was proved by Ahmad & Abeer [6]. The process keyword extraction (which includes; preprocessing, Morphology, Synonym Finding, Words Frequency, and Keyword Heuristic) was proved by the previous work. The other three processes will be explained as follows.

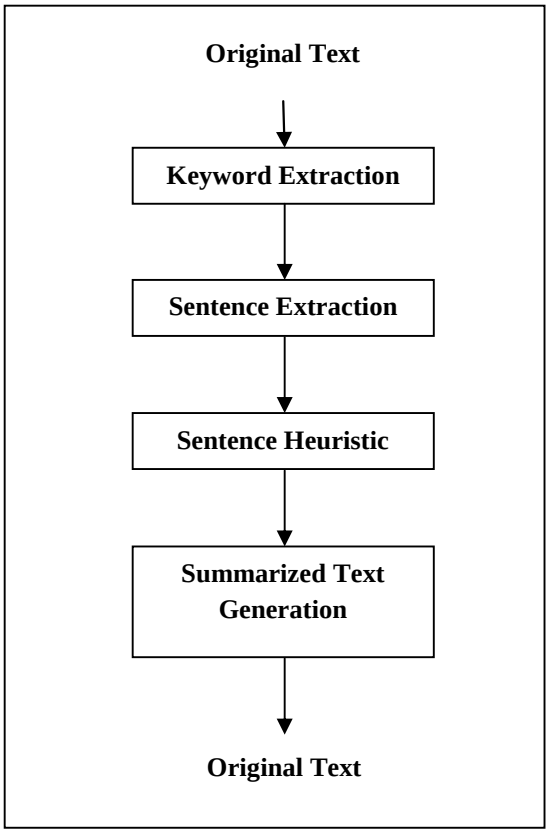

Figure (1) the proposed processes approach.

Sentence Extraction

To perform the process of sentence extraction, this system provides a pointer from each keyword to its sentence. The system stores more information about sentences; such as the order at the original text (sentence number), start token-position, end token-position, and Part-Of-Speech (POS). All sentences with keywords will be marked as important sentences, so important sentences are the results of this process.

Sentence Heuristic

During the previous process, information about important sentences is stored. Sentence information could be considered as heuristic values for the sentence. Now it will be used to compute one heuristic value for each important sentence. A pointer must be provided to link each heuristic value with its sentence. This heuristic value affected by; number of keywords at the sentence, the title, and POS as follows.

- Assign heuristic (H=0) for all important sentences.

- For each important sentence increase its heuristic by the number of keywords (No) inside the sentence (H=H + No).
- Increase sentence heuristic by 2 if the sentence is at title (H=H+2).
- For each important sentence, Verify POS if valid then increase its heuristic by one (H=H+1).

Summarized Text Generation

This process generates the summarized text from concatenating important sentences (not necessary all of them). This process chooses threshold value, to determine the amount of summary. For example, when threshold value=0.35 then, the summarized text includes only 0.35 from all sentences. Firstly, sentences heuristic are sorted in descending. Then choose part of original text depending on the threshold value. This process follows four steps to perform its task.

- Sort sentences heuristic in descending.
- Use the threshold value to determine the number of wanted sentences (No).
- From the important sentences, collect sentences that have the highest (No) heuristic values.
- Concatenating the collected sentences, keeping their order in original text.

EXPERIMENTAL RESULTS AND DISCUSSION

In order to test the proposed system, it is provided with a specific domain dictionary "artificial intelligence". The system is provided with a number of original texts, also provided with hand summary for each original text. To find the best threshold value, the proposed system is tested using different values 20%, 40%, and 60%. To compute the accuracy of this approach, the results are compared with the provided hand summary. So, the accuracy rate equal to rate of right sentences. Best accuracy was gained at threshold value 60% Figure (2). While worst accuracy was gained at threshold value 20%. Table (1) records each tested threshold value with its accuracy.

This approach depends on the keywords and other features that are extracted from the document to get the text summary as a result. It uses the EMA keyword extraction technique to determine the important statements. Features are recorded for each important statement such as; Number of keywords per sentence, first paragraph, and POS. They are used to compute a heuristic value. This gives the advance of finding the best sentences from the important ones. The system gives good results in comparison to manual summarization extraction. It can give the most compressed summary with high quality.

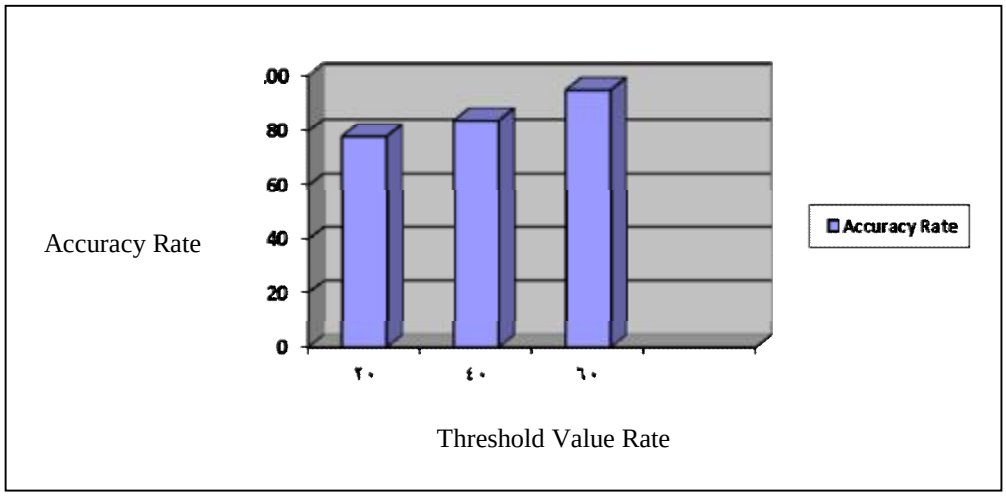

Figure (2) Accuracy rate of the proposed system.

Table (1) Results

Threshold Value	Accuracy Rate
20%	78%
40%	84%
60%	95%

CONCLUSIONS

1-The proposed approach uses EMA technique. It is a good tool to increase the performance of text summarization method with minimum dictionary size.

2-The best threshold value parameter in the proposed approach is 60%.

3-The proposed approach has the ability to gain good accuracy rate with minimum text summary. It gain 78% accuracy rate at 20% threshold value.

REFERENCES

[1].Vishal Gupta and Gurpreet Singh Lehal, "A Survey of Text Summarization Extractive Techniques", Journal Of Emerging Technologies In Web Intelligence, Vol. 2, No. 3, P. 258-268, August 2010.

[2].Dipanjan Das and Andr_e F.T. Martins, "A Survey on Automatic Text Summarization", Literature Survey for the Language and Statistics II course at CMU, November, 2007.http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.126.5100

[3].Elena Lloret, "Text Summarization: An Overview", (2008), Available At http://www.dlsi.ua.es/~elloret/publications/TextSummarization.pdf

[4]. Rafeeq Al-Hashemi,"Text Summarization Extraction System (TSES) Using Extracted Keywords", International Arab Journal of e-Technology, Vol. 1, No. 4, P. 164-168, June 2010.

[5]. Jasmeen Kaur and Vishal Gupta, "Effective Approaches For Extraction Of Keywords", IJCSI International Journal of Computer Science Issues, Vol. 7, Issue 6, P. 144-148, November 2010.

[6]. Ahmed T. Sadiq Al-Obaidi1 and Abeer Khalid AL-Mashhadany, "Using English Morphological Analyzer to Decrease the Dictionary Size ins Keywords Extraction Techniques", International Journal of Research and Reviews in Soft and Intelligent Computing (IJRRSIC), 2046-6412, Vol. 2, No. 1, P. 114-118, March 2012.

[7]. Ahmed Tariq Sadiq and Noor Amjed Hassan, "Learning- Based Text Summarization Approach Using Association Rules and Statistical Measurements", Iraqi Journal of Information Technology, Vol. 2, No. 3, 2008.

[8]. Ahmed Tariq Sadiq,Saran Akram Chawishly and Kanar Shukr Muhamad, "Text Summarization Using Hybrid Methods", Proceeding of the IAIT'09 Conference, Baghdad, Iraq, 2009.

[9]. Ahmed Tariq Sadiq and Enas Tariq Khuder, "English Text Summarization Using Statistical, Linguistics, Heuristics and Machine Learning Techniques", Proceeding of the 1st Computer Science Conference, Dep. Of Computer Science, University of Technology, Baghdad, Iraq, 2010.

[10]. Naresh Kumar Nagwani and Shrish Verma, "A Frequent Term and Semantic Similarity based Single Document Text Summarization Algorithm", International Journal of Computer Applications, Number 2 - Article 7, P. 36-40, 2011.