Awakening to reality

Available online at www.elixirpublishers.com (Elixir International Journal)

Computer Science and Engineering

Elixir Comp. Sci. & Engg. 60 (2013) 16203-16205

Elixir
ISSN: 2229-712X

Document summarization using differential evolution algorithm

Puspanjali Rout and Rasmita Rautray*
Department of CSE, ITER, SOA University.

ARTICLE INFO

Article history:
Received: 17 May 2013;
Received in revised form:
17 June 2013;
Accepted: 4 July 2013;

Keywords
Summarization,
Term frequency,
Sentence similarity,
Differential Evolution algorithm.

ABSTRACT

The use of document summarization allows a user to get a sense of the content of full document, or to know its information content without reading all sentences within the document. Data reduction helps user to find the required information quickly without having to waste time in reading the whole document. This paper presents a method to generate a summary from the original document. And the method includes several characteristics such as sentence-id, position of each term in a sentence, term frequency, sentence similarity measure and weight of each and every sentence. To solve the optimization problem differential evolution (DE) algorithm is used, which can choose the optimal summary. DE algorithm is based on a fitness function and selection of fitness function is crucial for the good performance of DE algorithm.

1. Introduction

Summarization is the process of reducing a text document in order to create a summary where, the number of lines of the summarized document is less than that of the original document. It retains the most important points of the original document. It is an abbreviated and accurate representation of a text that keeps the most essential contents present in it. It can be considered as a true substitute of the document, giving its outline in brief. The primary function of a summary is to indicate and predict the structure and content of the text thereby saving the user's time; without the need of actually going through the entire document. Automatic document summarization is an important research area in natural language processing (NLP). The technology of automatic document summarization is developing and may provide a solution to the information overload problem.

Document summarization tasks can be classified into single-document and multi-document summarization. In single document summarization, the summary of only one document is to be built, while multi-document summarization is aims at extracting major information from multiple documents [6, 8]. This paper describes single document summarization.

Xiaojuan Zhao et al. [2] proposed a new method of query-focused multi-document summarization based on genetic algorithm, genetic algorithm is used to extract the sentences to form a summary, and it is based on a fitness function formed by three factors. The proposed method can improve the performance of summary.

Yan-Xiang He et al. [1] proposed multi-document summarizer using genetic algorithm-based sentence extraction (MSBGA) regards summarization process as an optimization problem where the optimal summary is chosen among a set of summaries formed by the conjunction of the original articles sentences. To solve the NP hard optimization problem, MSBGA adopts genetic algorithm, which can choose the optimal summary on global aspect. The evaluation function employs four features according to the criteria of a good summary: satisfied length, high coverage, high informativeness and low redundancy. To improve the accuracy of term frequency, MSBGA employs a novel method TFS, which takes word sense into account while calculating term frequency.

Cristina Lopez-Pujalte et al. [4] proposed Order-Based Fitness Function for Genetic Algorithms Applied to Relevance Feedback. Recently there have been appearing new applications of genetic algorithms to information retrieval, most of them specifically to relevance feedback. The evolution of the possible solutions is guided by fitness functions that are designed as measures of the goodness of the solutions. These functions are naturally the key to achieving a reasonable improvement, and which function is chosen most distinguishes one experiment from another. In previous work, we found that, among the functions implemented in the literature, the ones that yield the best results are those that take into account not only when documents are retrieved, but also the order in which they are retrieved.

You Ouyang et al. [3] proposed a study on position information in document summarization. Position information has been proved to be very effective in document summarization, especially in generic summarization. Existing approaches mostly consider the information of sentence positions in a document, based on a sentence position hypothesis that the importance of a sentence decreases with its distance from the beginning of the document. In this paper, we consider another kind of position information, i.e., the word position information, which is based on the ordinal positions of word appearances instead of sentence positions. An extractive summarization model is proposed to provide an evaluation framework for the position information. The resulting systems are evaluated on various data sets to demonstrate the effectiveness of the position information in different summarization tasks.

This paper proposed a new method of document summarization based on differential evolution algorithm, DE is used to extract the sentence to form a summary and it is based on a fitness function formed by sentence similarity factor.

The rest of the paper is organized as follows. Section 2 details system overview. Section 3 details differential evolution algorithm. Section 4 describes the application of the proposed

Tele:
E-mail addresses: rasmitarautray@yahoo.co.in

algorithm on a sample test collection and evaluation result finally; we end this paper with conclusion.

2. System Overview

Our summarization system is design with the extractive framework. Important sentences are extracted and reorganized to form a summary. Thus the whole system is divided into three modules: preprocessing, processing and summary generation. The flowchart is as figure.1.

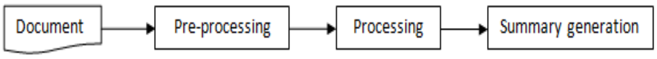


Figure 1: System Flowchart

2.1 Pre-processing

In pre-processing, many patterns are used to reduce one or few words from the original sentences without losing much information. Then, document is segmented into sentences.

2.2 Processing

In processing step, Differential evolution algorithm is used to extract summary sentences. Here sentences of a document are modeled as vectors [9] using vector space model. The execution of the differential evolution is similar to other evolutionary algorithms like genetic algorithms or evolution strategies. The evolutionary algorithms differ mainly in the representation of parameters (usually bi-nary strings are used for genetic algorithms while parameters are real-valued for evolution strategies and differential evolution) and in the evolutionary operators.

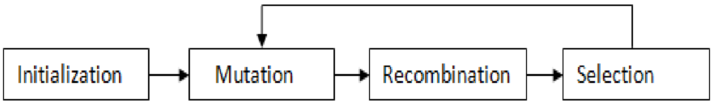


Figure 2: DE Algorithm Procedure

After sentence segmentation, an effective method is required to compute the similarity between sentences [5] Sentence similarity is calculated by most widely used vector space model (vsm). Here we have used Cosine similarity to measure similarity between two vectors of n-dimensions by finding the cosine of the angle between them. The method of tf-idf is easier to calculate the cosine of the angle between the vectors.

For sentences S_i =[p_1 ,p_2,…p_k] and S_j=[q_1 ,q_2…q_k], the sentence similarity is computed as:

$$Sim(S_i, S_j) = \frac{\sum_k p_i * q_j}{\sqrt{(\sum_k p_i^2) * (\sum_k q_j^2)}} \qquad (1)$$

2.3 Summary Generation

Our aim is to find a summary using DE algorithm. Here in this paper, initial population for DE algorithm is the term frequency matrix. Sentence extraction is an approach to sentence compression [10]. As compression rate decreases, the summary will be more concise [7]. A fitness function (f) is used to calculate the fitness of each chromosome and some control parameters are used like crossover probability (P_c) and mutation probability (P_m), where (P_c =0.8 and P_m=0.5) .

$$f = \frac{(\beta * wt + \gamma * sim)}{\beta + \gamma} \qquad (2)$$

Where β and γ are real numbers between 0 and 1, defined by the user.

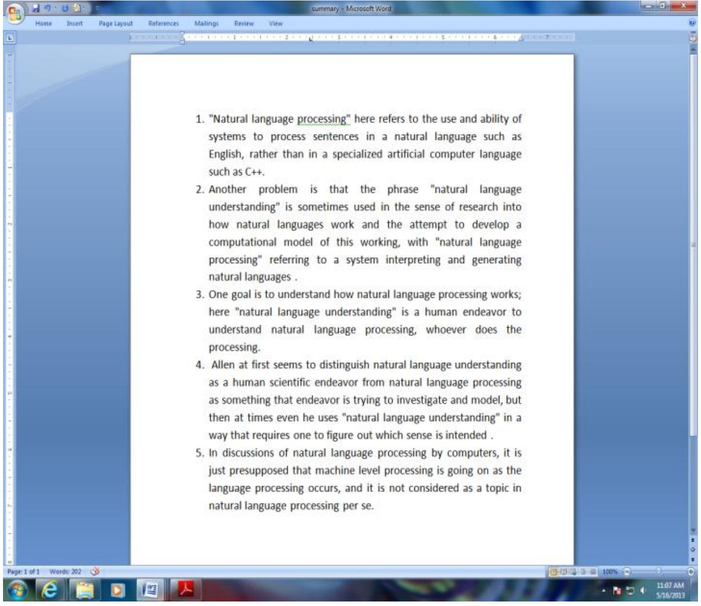
The stopping criteria of DE could be a given number of consecutive iterations within which no improvement of summary occurs.

3. Result Analysis:

Here we have generate a summary by taking a paragraph from our system and from an automatic summarizer and then compared it. In our system, to reduce the length of the summary we are taking a threshold value (depending upon the weight of each summary sentence).

If weight of the sentence is less than the threshold value, then it should be deleted from the summary. After comparing our summary with Automatic Summarizer and Microsoft word summarizer summary, our summary gives meaningful information as well as the better summary than the other summary generated by the Automatic summarizer and Microsoft word summarizer.

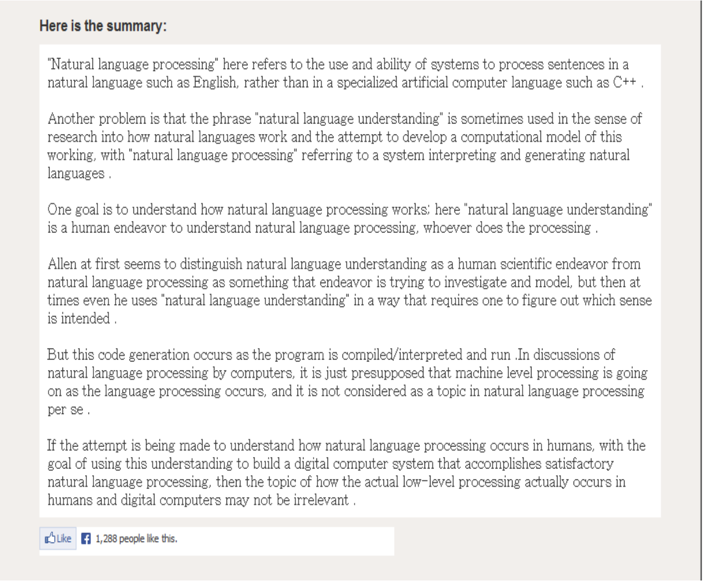
The summary generated by our proposed system is given below in fig. 3.



1. "Natural language processing" here refers to the use and ability of systems to process sentences in a natural language such as English, rather than in a specialized artificial computer language such as C++.
2. Another problem is that the phrase "natural language understanding" is sometimes used in the sense of research into how natural languages work and the attempt to develop a computational model of this working, with "natural language processing" referring to a system interpreting and generating natural languages .
3. One goal is to understand how natural language processing works; here "natural language understanding" is a human endeavor to understand natural language processing, whoever does the processing.
4. Allen at first seems to distinguish natural language understanding as a human scientific endeavor from natural language processing as something that endeavor is trying to investigate and model, but then at times even he uses "natural language understanding" in a way that requires one to figure out which sense is intended .
5. In discussions of natural language processing by computers, it is just presupposed that machine level processing is going on as the language processing occurs, and it is not considered as a topic in natural language processing per se.

Figure 3: Summary generated using DE algorithm

Summary generated according to the no of sentences entered in the summarizer is given below in fig.4.



Here is the summary:

"Natural language processing" here refers to the use and ability of systems to process sentences in a natural language such as English, rather than in a specialized artificial computer language such as C++ .

Another problem is that the phrase "natural language understanding" is sometimes used in the sense of research into how natural languages work and the attempt to develop a computational model of this working, with "natural language processing" referring to a system interpreting and generating natural languages .

One goal is to understand how natural language processing works; here "natural language understanding" is a human endeavor to understand natural language processing, whoever does the processing .

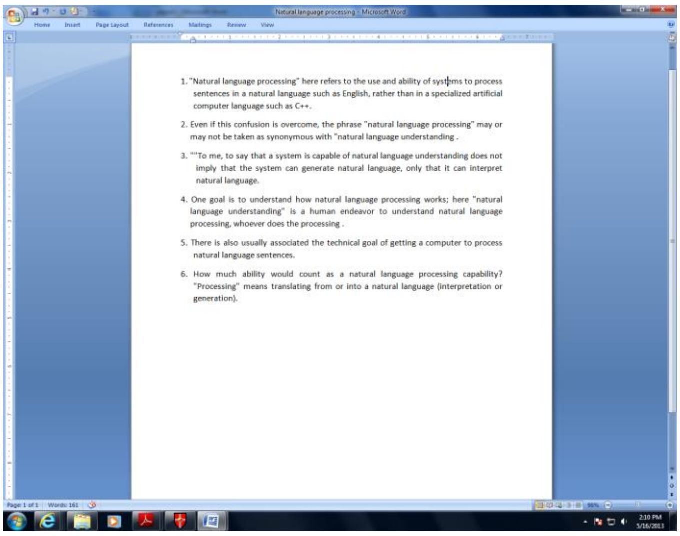
Allen at first seems to distinguish natural language understanding as a human scientific endeavor from natural language processing as something that endeavor is trying to investigate and model, but then at times even he uses "natural language understanding" in a way that requires one to figure out which sense is intended .

But this code generation occurs as the program is compiled/interpreted and run .In discussions of natural language processing by computers, it is just presupposed that machine level processing is going on as the language processing occurs, and it is not considered as a topic in natural language processing per se .

If the attempt is being made to understand how natural language processing occurs in humans, with the goal of using this understanding to build a digital computer system that accomplishes satisfactory natural language processing, then the topic of how the actual low-level processing actually occurs in humans and digital computers may not be irrelevant .

Figure 4: Summary generated by automatic summarizer

Summary generated according to the percentage of total sentence from the original document by using Microsoft word summarizer is given below in fig.5.



1. "Natural language processing" here refers to the use and ability of systems to process sentences in a natural language such as English, rather than in a specialized artificial computer language such as C++.
2. Even if this confusion is overcome, the phrase "natural language processing" may or may not be taken as synonymous with "natural language understanding .
3. ""To me, to say that a system is capable of natural language understanding does not imply that the system can generate natural language, only that it can interpret natural language.
4. One goal is to understand how natural language processing works; here "natural language understanding" is a human endeavor to understand natural language processing, whoever does the processing .
5. There is also usually associated the technical goal of getting a computer to process natural language sentences.
6. How much ability would count as a natural language processing capability? "Processing" means translating from or into a natural language (interpretation or generation).

Figure 5: Summary generated by Microsoft word summarizer

4. Conclusion

Evaluation result of Document Summarization using DE algorithm is an effective method. We take sentence similarity into account while designing the evaluation function for DE, which is helpful to improve the performance of summarization. We will work on multi-document summarization and other algorithms in our future work to improve the performance of summary.

References

[1] He Y., Liu D., Ji D., Yang H.; Teng C. "MSBGA: A Multi-Document Summarization System Based on Genetic Algorithm". *International Conference on Machine Learning and Cybernetics*, 2006, 2659 – 2664.

[2] Zhao X., Tang J. "Query-focused Summarization Based on Genetic Algorithm", *International Conference On Measuring Technology and Mechatronics Automation*, 2010, 968-971.

[3] Ouyang Y., Li W., Lu Q., Zhang R. "A Study on Position Information in Document Summarization", Coling 2010: Poster Volume, 2010, 919-927.

[4] Pujalte C., Guerrero-Bote V., Pujalte C., Guerrero-Bote V. "Order-Based Fitness Functions for Genetic Algorithms Applied to Relevance Feedback", *Journal Of The American Society For Information Science And Technology*, 2003,54(2):152-160.

[5] Kumar A., Premchand P., Govardhan A. "Query-Based Summarizer Based on Similarity of Sentences and Word Frequency", *International Journal of Data Mining & Knowledge Management Process (IJDKP)*, 2011,1(3).

[6] Kowsalya R., Priya R., Nithiya P. "Multi Document Extractive Summarization Based On Word Sequences". *International Journal of Computer Science*, 2011, 8(2):510-517.

[7] Alguliev R.,Aliguliyev R. "Evolutionary Algorithm for Extractive Text Summarization". *Intelligent Information Management*, 2009, 1(2):128-138**.**

[8] Lin C. and Hovy E. From Single to Multi-document Summarization: A Prototype System and its Evaluation. Proceedings of the 40th Annual Meeting of the *Association for Computational Linguistics* (ACL), Philadelphia, 2002.457-464.

[9] Ji X. "Research on the Automatic Summarization Model based on Genetic Algorithm and Mathematical Regression". *International Symposium on Electronic Commerce and Security*, 2008, 488-491.

[10] Nomoto T., "Discriminative sentence compression with conditional random fields". *Information Processing & Management*, 2007, 43(6):1571-1587.