Document Summarization Using Sentence Features

Rasmita Rautray, Department of Computer Science and Engineering, Institute of Technical Education & Research, Siksha ‘O’ Anusandhan University, Bhubaneswar, India

Rakesh Chandra Balabantaray, Department of Computer Science and Engineering, International Institute of Information Technology, Bhubaneswar, India

Anisha Bhardwaj, Department of Computer Science and Engineering, Institute of Technical Education & Research, Siksha ‘O’ Anusandhan University, Bhubaneswar, India

ABSTRACT

Problem of exponential growth of information available electronically, there is an increasing demand for text summarization. Text summarization is the process of extracting the contents of the original text in a shorter form that provides useful information to the user. This paper presents a summarizer to produce summaries while reducing the redundant information and maximizing the summary relevancy. The proposed model takes several features into an account, including title feature, sentence weight, term weight, sentence position, inter sentence similarity, proper noun, thematic word and numerical data. The score of each feature for the model can be obtained from the document sets. However, the results of such models are evaluated to measure their performance based on F-score of extracted sentences at 20% compression rate on a C-50 data corpus. Experimental studies on C-50 data corpus, PSO summarizer show significantly better performance compared to other summarizer.

Keywords: Summarization, Sentence feature, Differential Evolution, Particle Swarm Optimization, FLANN

With the exponential growth of textual information available electronically, it makes more difficult to access the usable information (Alguliev et al., 2012; Mendoza et al., 2014). Therefore, obtaining the desired information within a short amount of time becomes a serious problem in information age. As such, new technologies that can process information necessary for end user, which can access in summary form & without losing the most important aspects therein. A summary is the main objective of summarization method, which can be categorized into two ways- extraction and abstraction (Alguliev & Aliguliyev, 2009; Ferreira. et al., 2014). Extraction involves sentence extraction from the corpus and chronological concatenation to form a summary, where as abstraction involves generating novel sentences from the information extracted by the corpus. But, the summary has no specific classification, though its categorization is based on different criteria. Depending on the size of the document to be summarized the summary can be a single document or multi document summary (Rout & Rautray, 2013). In a single document summary the aim of summarization is to produce a

DOI: 10.4018/IJIRR.2015010103
concise summary from a single document and, if it produces a concise summary from related multiple documents, it is known as multi-document summary. Also, depending on the target audience, summaries may be generic, query-based (Steinberger & Ježek, 2012). Generic summaries do not depend on the audience for whom the summary is intended. Query-based summaries respond to the query made by the user. In addition, depending on the language of the document, they may be monolingual or multilingual, and regarding document genre may be scientific article, news, blogs, and so on.

Due to the problem of improvising the performance of text selection in document summarization using statistical tools, a number of global optimization techniques and learning techniques have been proposed in literature. In the use of global optimization technique, the inputted text are optimized to obtained optimal summary, whereas learning techniques that rely on pre-existing document-summary pairs, called supervised techniques and based on properties and heuristics derived from the text, called unsupervised techniques. Supervised extractive summarization techniques treat the summarization task as a two-class classification problem at the sentence level (Fattah & Fuji, 2009; Nandhini & Balasundaram, 2014). But, many unsupervised methods have been developed for document summarization by exploiting different features and relationships of the sentences.

This paper presents a corpus based generic monolingual single document summarization system. The system is based on the statistical features of each sentence to obtain coverage of the summary and cohesion of summary sentences. The rest of the paper is organized as follows: Section-2 presents a complete overview of the summarization system. The model section of the system uses differential evolution algorithm, particle swarm optimization algorithm and functional link artificial neural network algorithm. While the evaluation using data sets, along with a comparison and analysis are presented in Section 3; and finally, Section 4 presents the conclusions and future work.

OVERVIEW OF SUMMARIZATION SYSTEM

Figure 1 illustrates the proposed automatic model for summarization. It includes three basic steps to generate summary which are pre-processing, feature extraction and summary generation.

Preprocessing

Initially the document is segmented into sentences and words for each sentence are extracted. Then the functional words or stop words like “a”, “the”, “of” (frequently occurring insignificant words) are removed from the word list. The words remaining in the sentences are stemmed.

Feature Extraction

Feature is one of the important aspects of any text mining. Therefore the following features for each sentence need to be prepared for input to the optimization model (Table 1).

1. \textbf{ft1}= Title Feature: It is the similarity between this sentence & the document title. The score of ft1 is calculated as follows:

\[
Score_{ft1}(s) = \frac{|KWDS \cap KWDT|}{|KWDS \cup KWDT|}
\]

2. \textbf{ft2}= Sentence Length: This feature is employed to penalize sentences that are too short, since these sentences are not expected to belong to the summary. We use longest sentence length of the sentence for normalization.

\[
Score_{ft2}(s) = \frac{SL(i)}{LSL}
\]  

3. \textbf{ft3}= Average Sentence Weight: This feature specifies the weight of each sentence by taking term frequency into an account. The score of ft3 is calculated as follows: