International Journal of Innovative Computing, Information and Control
Volume **5**, Number **7**, July **2009**

 ISSN 1349-4198
pp. **1987–1996**

AUTOMATIC TEXT SUMMARIZATION USING SUPPORT VECTOR MACHINE

Nadira Begum[1], Mohamed Abdel Fattah[1,2] and Fuji Ren[1,3]

[1]Faculty of Engineering
University of Tokushima
2-1 Minamijosanjima
Tokushima 770-8506, Japan
{ begum; mohafi; ren }@is.tokushima-u.ac.jp

[2]FIE, Helwan University
Cairo, Egypt

[3]School of Information Engineering
Beijing University of Posts and Telecommunications
Beijing 100088, P. R. China

Received June 2008; revised October 2008

Abstract. *This work investigates different text features to select the best one and proposes an approach to address automatic text summarization. This approach is a trainable summarizer, which takes into account several features, including sentence position, sentence centrality, sentence resemblance to the title, sentence inclusion of name entity, sentence inclusion of numerical data, sentence relative length, Bushy path of the sentence and aggregated similarity for each sentence to generate summaries. First we investigate the effect of each sentence feature on the summarization task. Then we use all features score function to train Support Vector Machine (SVM) in order to construct a text summarizer model. The proposed approach performance is measured at several compression rates (CR) on a data corpus composed of 100 English articles from the domain of politics.*
Keywords: Automatic summarization, Support vector machine, Text features

1. **Introduction.** With the huge amount of information available electronically, there is an increasing demand for automatic text summarization systems. Text summarization is the process of automatically creating a compressed version of a given text that provides useful information for the user. Text summarization addresses both the problem of selecting the most important portions of text and the problem of generating coherent summaries. There are two types of summarization: extractive and abstractive. Extractive summarization methods simplify the problem of summarization into the problem of selecting a representative subset of the sentences in the original documents. Abstractive summarization may compose novel sentences, unseen in the original sources. However, abstractive approaches require deep natural language processing such as semantic representation, inference and natural language generation, which have yet to reach a mature stage nowadays.

The process of text summarization can be decomposed into three phases: analysis, transformation, and synthesis. The analysis phase analyzes the input text and selects a few salient features. The transformation phase transforms the results of analysis into a summary representation. Finally, the synthesis phase takes the summary representation, and produces an appropriate summary corresponding to users' needs. In the overall process, compression rate, which is defined as the ratio between the length of the summary