An Improved Feature-Based Automatic Text Summarization System

By

Madhi Ahmed Ali

Supervisor
Prof. Dr. Ali A. Al-Dahoud

Co-Supervisor
Dr. Bilal H. Hawashin

Abstract

In this work, we propose an efficient text summarization system that uses a combination of existing and improved sentence features to extract the summary under supervised learning. Many works in the literature used sentence features, but some of these features suffer from weakness issues. Improving such features would contribute in improving the performance of text summarization systems. For this purpose, the study proposes improvements to Sentence position, Sentence length, and Key word sentence features. Afterwards, the system find the optimal combination between these features and some existing features such as Term frequency, Sentence centrality, Title similarity, and Upper case of word. By using machine learning techniques, mainly Support Vector Machines (SVM), Naive Bayes (NB), and Decision Tree classifiers, this thesis evaluates two feature groups: a combination of seven features without any improvements, and the same seven features after making some improvements on Sentence position, Sentence length, and Key word sentence features. Proposed system consists of three main stages which are: preprocessing, processing, and post processing. Processing which is the most

important stage contains feature extraction and machine learning steps. The study includes two sets of data; each one consists of 100 English documents, which are used as training and testing data by Machine Learning step to evaluate the accuracy. The evaluation metrics includes Precision, Recall and F-measure.

Experimental results show that improving some features contributes in improving the performance of the text summarization system. A new system increasing the accuracy for J48, SVM, and N.B by 3.5%, 4.1%, and 3.5% respectively for 40% Compression Rate (CR); in contrast to, 3.1%, 3.2%, and 3.3% with 20% CR.