Automatic Headline Generation for Arabic Textual Documents

Salah Gad Mohamed Foda ,Fahad Alotaiby; Ibrahim Alkharashi

Abstract

A headline is considered a condensed summary of a document. The necessity for automatic headline generation has been on the rise due to the need to handle a huge number of documents, which is a tedious and time-consuming process. Instead of reading every document, the headline can be used to decide which ones contain important and relevant information. There are two major approaches to automatic headline generation. The first is linguistic, in which the knowledge about the structure of the language itself is considered. The second approach is statistical and it comprises all quantitative approaches to automated language processing. However, the Arabic language has a different statistical structure than the English language, and requires special treatment to generate Arabic headlines, especially when there is no dedicated technique for the Arabic language. Therefore, two new statistical methods in automatic headline generation have been developed to create representative headlines for textual documents in the Arabic language. The first is an extractive method based on character cross-correlation, and the second one is an abstractive method based on the hidden Markov model (HMM). The extractive method achieved ROUGE-L of (0.1938) and the HMM method achieved ROUGE-L of (0.2332). In addition, both techniques were assessed via human examiners who evaluated the resulting headlines.

KACST 2012, February