Jordan University of Science and Technology

Automatic Keyphrase Extraction for Arabic News Documents based on KEA System

Authors: Rehab Duwairi, Mona Hedaya

Abstract: A keyphrase is a sequence of words that play an important role in the identification of the topics that are embedded in a given document. Keyphrase extraction is a process which extracts such phrases. This has many important applications such as document indexing, document retrieval, search engines, and document summarization. This paper presents a framework for extracting keyphrases from Arabic news documents which is based on the KEA system. It relies on supervised learning, Na?ve Bayes in particular, to extract keyphrases. Two probabilities are computed: the probability of being a keyphrase and the probability of not being a keyphrase. The final set of keyphrases is chosen from the set of phrases that have high probabilities of being keyphrases. The novel contributions of the current work are that it provides insights on keyphrase extraction for news documents written in Arabic. It also presents an annotated dataset that was used in the experimentation. Finally, it uses Na?ve Bayes as a medium for extracting keyphrases.