Abstract

In this work the improvement of automatic keyphrases extraction using deep linguistic features and supervised machine learning algorithm is discussed. The n-gram method for extracting important keyphrases produces huge number of candidate terms. Many of those terms are non-keyphrases either because they are linguistically non expressive terms or due to redundancy in sense. The objective is to restrict the number of candidate terms and keeping the relevant ones. This work is an extension to a previous one in keyphrase extraction for Arabic documents. The proposed work covers the deep linguistic features of the candidate terms. To capture the well-structured terms a new-added definite structure feature is introduced and tested. A set of linguistic features of the previously assigned candidate terms are applied to a supervised machine learning technique to classify the candidates as keyphrases or not. The experiments carried out showed that the proposed technique improves the accuracy of extracting keyphrases relative to the previous version and other available extractors.

Index Terms

Computer Science                  Pattern Recognition
Keywords

Keyphrase Extraction, Arabic Keyphrases, Information Retrieval, Classification Methods, Computational Linguistics.