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.
2. CBA. Data mining tool Downloading page at:
   Conference on Arabic Language Resources and Tools. The MEDAR Consortium, Cairo, Egypt.
4. El-Shishtawy, T., El-Ghannam, F. An Accurate Arabic Root-Based Lemmatizer for
   Volume 9.
   8th International Conference on Informatics and Systems INFOS, Egypt.
   Domain-specific keyphrase extraction. Proceedings of the Sixteenth International Joint
   Conference on Artificial Intelligence (IJCAI-99), pp. 668-673. California: Morgan Kaufmann.
7. G¨onen, E. Automated text summarization and keyphrase extraction. 2006. Master thesis,
   Bilkent University.
   Proceedings of the Association for Computational Linguistics (ACL), Baltimore, Maryland:
   Association for Computational Linguistics.
   Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing,
   Sapporo, Japan.
12. Turney, D. 2000. Learning algorithms for keyphrase extraction. Information Retrieval, 2,
    pp.303-336.
15. Witten, I., frank, E. 2000. Data Mining: practical machine learning tools and techniques
    with Java implementations, Morgan Kaufmann , San Francisco.
    Press.
17. Witten, I., Paynter, W., Frank, E.,Gutwin, C., and Nevill-Manning, G. 2000. KEA:
    Practical Automatic Keyphrase Extraction. Working Paper 00/5, Department of Computer
    Science. The University of Waikato.

**Index Terms**

Computer Science  
Pattern Recognition
Keywords

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