Tarek Kanan and Edward A. Fox. ”Automated Arabic Text Classification with P-Stemmer, Machine Learning, and a Tailored News Article Taxonomy”. Journal of the Association for Information Science and Technology, Impact Factor 2.28. Willy Publisher, USA. DOI: 10.1002/asi.23609

Abstract:

Arabic news articles in electronic collections are difficult to study. Browsing by category is rarely supported. Although helpful machine-learning methods have been applied successfully to similar situations for English news articles, limited research has been completed to yield suitable solutions for Arabic news. In connection with a Qatar National Research Fund (QNRF)-funded project to build digital library community and infrastructure in Qatar, we developed software for browsing a collection of about 237,000 Arabic news articles, which should be applicable to other Arabic news collections. We designed a simple taxonomy for Arabic news stories that is suitable for the needs of Qatar and other nations, is compatible with the subject codes of the International Press Telecommunications Council, and was enhanced with the aid of a librarian expert as well as five Arabic-speaking volunteers. We developed tailored stemming (i.e., a new Arabic light stemmer called P-Stemmer) and automatic classification methods (the best being binary Support Vector Machines classifiers) to work with the taxonomy. Using evaluation techniques commonly used in the information retrieval community, including 10-fold cross-validation and the Wilcoxon signed-rank test, we showed that our approach to stemming and classification is superior to state-of-the-art techniques.

Keywords: Natural language processing; Information retrieval; Digital libraries