Abstract

Latin-based languages work smoothly within the traditional text mining techniques due to being definite and the natural limited alternatives of words meanings. On the other hand, in the Arabic language, we are facing 2 main differences: 1) the way, the Arabic language is being written today without diacritics in 99% of the text will make the text interpretation at the level of two consecutive words and even in some cases at the level of sentences indefinite 2) even with diacritics, Arabic words are very loose; each word in Arabic may bear more than one meaning regarding the context. Hence handling text in Arabic in the same manner that Latin languages do, will be rather time wasting. We need to rely on different techniques in order to enrich the criteria which will be adopted in text analysis. We propose a domain-specific approach that yielded excellent results with some of Arabic text analysis aspects. Several classifiers have been built and tested for this purpose. This approach was compared to others that don’t use the domain-specific approach; the paper concludes that the results obtained from the adopted technique are more appealing and promising.

References

- Open Directory RDF Dump, retrieved from: http://rdf.dmoz.org/
- David F. Prenatt, Jr., Life after the Open Directory Project, Traffick.com (June 1, 2000).
- An empirical study of the domain dependence of supervised word sense disambiguation systems, 2000 Article. Bibliometrics Data Bibliometrics.

Index Terms

Computer Science Information Sciences

Keywords