Statistical/Corpus Based Methods for Improved Bilingual (Arabic/English) Web Search

Adnan Yahya, Anwar Hithnawi, Ali Salhi and Mearna Fawadleh
Computer Systems Engineering Department
Birzeit University, Palestine

Abstract

With the vast growth of the Arabic text documents stored on Web and data warehouses on the Internet, the need for powerful natural language processing methods to help people finding and managing Arabic information in an efficient is becoming more acute.

Through this research we are targeting the design and implementation of text mining tools that are able to efficiently index, process, search, and categorize Arabic data in large quantities in preparation for future searches.

For that we needed to deal with the challenges Arabic introduces for natural language processing and information retrieval, automatic Arabic document categorization, and Arabic query correction/suggestion and expansion. We focus on the issue of how to employ statistical/corpus based natural language processing methods to guarantee the best way of managing and finding Arabic information and the development of the corpus.