A Preliminary Study for Building an Arabic Corpus of Pair Questions-Texts from the Web: Aqa-Webcorp

Authors : Wided Bakari, Patrce Bellot, Mahmoud Neji

Abstract : With the development of electronic media and the heterogeneity of Arabic data on the Web, the idea of building a clean corpus for certain applications of natural language processing, including machine translation, information retrieval, question answer, become more and more pressing. In this manuscript, we seek to create and develop our own corpus of pair's questions-texts. This constitution then will provide a better base for our experimentation step. Thus, we try to model this constitution by a method for Arabic insofar as it recovers texts from the web that could prove to be answers to our factual questions. To do this, we had to develop a java script that can extract from a given query a list of html pages. Then clean these pages to the extent of having a database of texts and a corpus of pair's question-texts. In addition, we give preliminary results of our proposal method. Some investigations for the construction of Arabic corpus are also presented in this document.

Keywords : Arabic, web, corpus, search engine, URL, question, corpus building, script, Google, html, txt

Conference Title : ICMLA 2016 : International Conference on Machine Learning and Applications

Conference Location : Copenhagen, Denmark

Conference Dates : June 27-28, 2016