Jordan University of Science and Technology

Using Web Mining Techniques to Build a Multi-Dialect Lexicon of Arabic

Authors: Rania Al-Sabbagh, Roxana Girju, Mark Hasegawa-Johnson, Abbas Benmamoun, Rehab Duwairi, Eiman Mustafawi

Abstract: This paper presents an automatic technique for building a multi-dialect lexicon of four Arabic dialects, namely Egyptian Arabic (EA), Gulf Arabic (GA), Iraqi Arabic (IA) and Moroccan Arabic (MA). Each Modern Standard Arabic (MSA) entry is to be mapped to its synonyms in the four dialects on the basis of the correlations among their word co-occurrence patterns. The main obstacle, however, for building such a lexicon automatically is the lack of parallel corpora of different Arabic dialects and the scarceness of Arabic dialect corpora in general which are necessary for acquiring statistically reliable word co-occurrences. In order overcome such an obstacle, a circular ? rather than a parallel ? acquisition technique is to be used. According to the circular acquisition technique, the acquisition of word co-occurrences of one dialect is conditioned by the word co-occurrences acquired for the other dialects. That is, word co-occurrences of the first dialect are validated as possible word co-occurrences of the second dialects and word co-occurrences of both dialects are validated for the third dialect and so on and so forth. This technique manages to overcome the lack of parallel corpora for Arabic dialects since it can work on unrelated Web documents which are more frequently available than parallel corpora. Moreover, this technique is to handle some limitations of current search engines including search result duplications and the restriction on giving 1000 search results as a maximum. Despite the apparent contributions of the proposed technique and the promising results being achieved, it does not come without question. The crucial question to the proposed technique is about the direction of the acquisition process (i.e. how dialects are to be arranged in the circle of word co-occurrences acquisition). Since the proposed technique is mainly to enable digging deeper in the Web content of the scarce dialects, the authors assumed that they should star