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

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 start with dialects with large Web content so as to acquire as many possible collocations as possible and move to the dialects of small Web content to check the possibility of these collocations in them. The size of the Web content of each dialect is to be estimated using the methodology followed by Kilgarriff and Grefenstette (2003).

Preliminary results for the proposed technique and methodologies are promising, especially that mapping dialectical synonymous words to their MSA synonyms enables transferring the Part-of-Speech (POS) tag(s) and the semantic features of gender and number of the MSA word to its dialectical synonyms. This information is indispensible for many NLP applications and tasks including, but not limited to, POS tagging, parsing, anaphora resolution, Machine Translation, Text Summarization among others. Moreover, through their MSA synonyms, dialectical words get access to MSA available NLP tools and resources which are considerably available.

The paper can be extended in many directions. First, only content words are considered here; yet function words are expected to be more complicated because they are not to be defined in terms co-occurring words but in terms of their grammatical behavior. Second, synonyms across different dialects are not necessarily one-to-one alignments. Finally, inter-dialectical words are still beyond the scope of this paper and thus they should be considered in further research.