Computer Processing of Arabic Script-based Languages: Current State and Future Directions

Ali Farghaly
SYSTRAN Software, Inc.
9333 Genesee Ave
San Diego, CA 92121, USA
alifarghaly@aol.com

Arabic script-based languages do not belong to a single language family, and therefore exhibit different linguistic properties. To name just a few: Arabic is primarily a VSO language whereas Farsi is an SVO and Urdu is an SOV language. Both Farsi and Urdu have light verbs whereas Arabic does not. Urdu and Arabic have grammatical gender while Farsi does not. There are, however, linguistic and non-linguistic factors that bring these languages together. On the linguistic side it is the use of the Arabic script, the right to left direction, the absence of characters representing short vowels and the complex word structure. Non-linguistic common properties that bind the majority of speakers of these languages include: the Qur'an that every Moslem has to recite in Arabic, proximity of the countries speaking these languages, common history and, to a large extent, a common culture and historical influx . It is not surprising, then, that the surge of interest in the study of these languages and the sudden availability for funding to support the development of computational applications to process data in these languages come for all these languages at the same time.

This also occurs at crucial period in the field of Natural Language Processing (NLP). It is becoming increasingly evident that statistical and corpus-based approaches, though necessary, are not sufficient to address all issues involved in building viable applications in NLP. Arabic script-based languages share in different degrees an explosion of homograph and word sense ambiguity. The absence of the representation of short vowels in normal texts dramatically increases the number of ambiguities. At SYSTRAN, the average number of ambiguities of a token in many languages is 2.3, whereas in Modern Standard Arabic, it reaches 19.2. Dealing with such a problem represents a real challenge to NLP systems. Resolving ambiguity in NLP requires representation not only of linguistic and contextual knowledge but also of domain and world knowledge. It is not clear how number crunching of linguistic data could address this problem. Ambiguity in Arabic is enormous at every level: lexical, morphological and syntactic. Another serious problem is tokenization. It is extremely common in Arabic to find a token such as **"ورأيتهـــم"** which is actually a sentence consisting of a conjunction, a verb, a subject, an object in that order. Moreover, within the verb itself, there is tense, number and gender and mood. Within the object, which is only two alphabet letters, there is number, gender and case. The complexity of tokens and the abstractness of information, such as the meanings of prosodic templates (McCarthy, 1981), present challenges in the processing of Arabic script—based languages.

There has been steady progress in computational processing of Arabic script-based languages in the last few years. The greatest leap since the pioneering efforts made in the early 1980s in Arabic computational linguistics (Hlal, 1985;Ali 1985, 1987, 1988; Geith 1988; Farghaly, 1987), is the availability of Buckwalter's morphological analyzer and dictionary which has recently given a boost in that area. The great work at the LDC in the creation of a corpus of written and spoken Arabic as well as the Arabic tree bank is another important resource to the practitioners in the field. What is urgently needed in future research is work on syntactic analysis and ambiguity resolution.