AUTOMATIC EXTRACTION OF PREPOSITIONS IN A CORPUS OF MODERN STANDARD ARABIC WRITTEN TEXTS

Giuliano Lancioni

INTRODUCTION

An experiment in Natural Language Processing [NLP] of Modern Standard Arabic [MSA] written texts is presented, which aims to uniquely assign broad grammatical categories to words in a corpus by exploiting paradigmatic regularities in the lexicon, without any previous morphological or syntactic knowledge.

As usual in Arabic, the texts included in the corpus are unvocalized (i.e., basically, only consonants and long vowels are marked). Of course, this makes automatic reading and understanding much more difficult.

The approach taken capitalizes on a set of simple formal tests, inspired by strategies proposed in Classical Arabic linguistic thinking, which help to detect nominal or verbal forms according to the possibility to place some grammatical elements before them.

The task is subdivided in five subtasks; the core subtask (and the most difficult one) is the identification of simple one-letter prepositions, which in Arabic script are written together with the following word. The idea behind the identification of simple inseparable prepositions is to determine if the rest of the graphical word (i.e., the aggregate of preposition and following nominal) can be regarded as a “candidate nominal”, while the whole form cannot.¹

The system succeeds in assigning a grammatical category to a large subset (> 85%) of both data sets (training and testing, with a small degradation in performance in the latter case). Comparing the results with (quasi-)native speakers’ judgments shows a satisfactory level of accuracy (> 60% for recognized word types).

¹ The category “nominal” includes both nouns and adjectives, which are not easily distinguishable in MSA. For a discussion of categories in the Arabic linguistic tradition, see § 4.2 below.
In many NLP systems, the extraction of prepositions from written texts is a relatively easy task. In a language such as English, prepositions usually have the status of whole written words, and identification problems only come from homographs (for instance, after as a preposition and after as an adverb).

On the other hand, most simple prepositions in MSA are not graphically distinct from the first word of the following noun phrase: they are monosyllabic morphemes attached, both in writing and in pronunciation, to the following word: for instance, bi- ‘at, with’, li- ‘to’, ka- ‘as, like’. In the writing system of Arabic, these morphemes appear as a single consonant grapheme attached to the next word. This feature of graphic (and phonological) coalescence is mirrored by the identical term used by the Arabic grammatical tradition to refer to ‘(consonant) grapheme’ and ‘preposition’: ḥarf, literally ‘segment’.

This article is the report of an experiment of automatic extraction of prepositions from a significantly large subset of a corpus of written MSA texts taken from the Internet edition of Asharq Alawsat (aš-Šarq al-ʾ Awsat). The sample used in the experiment includes 32 daily issues of the newspaper, with about 1,100,000 words.

The main strategy adopted in the experiment is based upon the paradigmatic selection of word types with and without a preposition. This strategy has already been successfully employed in the identification of nominals—which are not easily distinguished from verbs in an unvocalized Arabic text—and has the advantage of allowing a high degree of accuracy without any human intervention.⁵ Though simpler, the philosophy of this experiment is analogous to statistical part-of-speech taggers, like Tnt (Brants 2000) and MXPOST (Ratnaparkhi 1996).

---

⁵ A parallel but different strategy has revealed itself very successful in the identification of specialized lexicon: by exploiting the classification of articles in the different sections of the newspaper (see Table 3 below), the ratio between occurrences of an item in a given section (say, economy or sports) and the total occurrences of the same item in the corpus allows one to mark items as (relatively) specialized or general. Thus, an item such as ka’s ‘cup’, which appears 200 times in the sports section of the newspaper out of 208 total occurrences (or 96.2%), shows up as a specialized, sport-related word. This is an interesting result, since ka’s is recorded in dictionaries as a general word, with a wider meaning of ‘holder for liquids, glass, cup’, but it has clearly become a predominantly sport-related word in MSA newspaper texts.