Journal of King Saud University – Computer and Information Sciences 29 (2017) 171–178

Contents lists available at ScienceDirect

Journal of King Saud University – Computer and Information Sciences

journal homepage: www.sciencedirect.com

Towards a standard Part of Speech tagset for the Arabic language

Imad Zeroual [a,*], Abdelhak Lakhouaja [a], Rachid Belahbib [b]

[a] Department of Mathematics and Computer, Science Faculty of Sciences, University Mohamed First, B-P 717, Oujda 60000, Morocco
[b] Doha Historical Dictionary of the Arabic Language, Doha, Qatar

ARTICLE INFO

Article history:
Received 13 April 2016
Revised 13 January 2017
Accepted 23 January 2017
Available online 2 February 2017

Keywords:
Natural Language Processing
Part of Speech
Tagging
Arabic tagset
TreeTagger

ABSTRACT

Part of Speech (PoS) tagging is still not very well investigated with respect to the Arabic language. Determining the PoS tags of a word in a particular context is difficult, primarily because there is no use of diacritics in most of contemporary texts. Consequently, the same word may be spelled in different ways. Further, detecting the difference between Arabic derivatives represents a very challenging issue for the majority of PoS taggers. Hence, the task of tagging the correct PoS tags requires advanced processing and the use of considerable resources. This study aims to design detailed hierarchical levels of the Arabic tagset categories and their relationships. These hierarchical levels allow easier expansion when required and produce more accurate and precise results. They are based on a comparative study and important references in Arabic grammar; they are also validated by experts in this field. In addition, the proposed tagset is implemented in a PoS tagger and tested via various experiments. We believe that our study makes a significant contribution to the literature because this work is an advancement in the direction of achieving a standard, rich, and comprehensive tagset for Arabic.

1. Introduction

Part of Speech (PoS) tagging is an important research area and the basis for a number of Natural Language Processing (NLP) tasks. Unfortunately, there is no standard PoS tagset used for Arabic Language Processing (ALP). In fact, only a small number of researchers are interested in the question of standards, especially in ALP. Consequently, it is difficult to benefit from existing PoS taggers or compare and evaluate different tagging approaches under the same conditions. Yet, several researchers have proposed tagsets that comply with their suitable objectives without considering Arabic grammatical features.

The majority of currently used tagsets are derived from English, which is a drawback for a morphologically complex language such as Arabic. The adaptation of such tagsets is problematic for Semitic languages as Zitouni (2014) claimed. "Approaches to PoS tagging were limited to English, resources for other languages tend to use 'tag sets', or inventories of categories that are minor modifications of the Standard English set". Moreover, the most widely used tagsets as standard guidelines, namely those recommended by the Expert Advisory Group on Language Engineering Standards (EAGLES), are designed for Indo-European languages. These guidelines are not entirely suitable for Arabic. Further, several of the current systems tend to target a PoS tagset that is not sufficiently suitable for different applications (Habash et al., 2009) (e.g., (Khoja, 2001; Darwish, 2002; Diab, 2007)).

The main challenge involved in constructing any NLP system for Arabic is amplified by the lack of language resources such as tagged corpora, which are fundamental for research and development in statistical computational linguistics (Farghaly and Shaalan, 2009). PoS tagging is one of the first processes that directly reflects the performance of other subsequent text processing (Albared et al., 2011). Habash and Sadat (2006) studied the effects of PoS tagging and demonstrated a positive influence on the quality of statistical machine translation.

Before addressing the PoS tagging process, the first requirement for the annotation of Arabic text is the compilation of a tagset that can accurately describe and address all the information regarding the language (Khoja et al., 2001). Further, an investigation of PoS tagging for Arabic indicates that using a complex tagset and then converting the resulting annotation to a smaller tagset provides a higher accuracy than tagging using the smaller tagset directly (Kübler and Mohamed, 2012).

* Corresponding author.
E-mail address: mr.imadine@gmail.com (I. Zeroual).
Peer review under responsibility of King Saud University.

Production and hosting by Elsevier

http://dx.doi.org/10.1016/j.jksuci.2017.01.006
1319-1578/

The present paper aims to develop the finest possible PoS tagset for Arabic and to produce more accurate and precise results that can be used to maximize the performance of subsequent ALP tasks such as syntactic parsing.

The proposed tagset is tested using a probabilistic tagging method. This method estimates the transition probabilities using a decision tree, which differs from other probabilistic taggers. Based on this method, a language-independent PoS tagger called TreeTagger is then adapted to use this tagset.

This paper is organized as follows: in Section 2, we provide background information regarding PoS tagging and specific approaches that attempt to solve the problem of PoS ambiguities. Furthermore, some popular PoS taggers for Arabic are presented. We illustrate the relevant existing tagsets with their drawbacks in Section 3. In Section 4, we describe the proposed tagset based on standard design criteria and compare it to similar projects. In Section 5, we present the usability test of the tagset via various experiments and discuss the findings. We conclude this paper in Section 6.

2. Background information

After providing a definition of PoS tagging, various approaches that have been adopted for this process are presented. Further, examples of Arabic PoS taggers are cited with their performance results.

2.1. Definition of Part of Speech tagging

PoS tagging is the ability to computationally determine what PoS tag of a word is activated by its use in a particular context (Albared et al., 2011). It is the task that involves managing ambiguity in processed text.

2.2. Tagging approaches

The task of identifying all the possible PoS tags of a word is not difficult, thanks to the existence of efficient Morphosyntactic analysers for Arabic words such as "AlKhalil Morpho Sys" (Boudchiche et al., 2017), Madamira (Pasha et al., 2014), and (Buckwalter, 2004). However, it remains difficult to achieve disambiguation.

In earlier interesting works by other researchers (Farghaly and Shaalan, 2009; Maamouri and Bies, 2010), the reason why ambiguity exists on numerous levels in Arabic is presented. For example, analysing the Arabic word " ثمن "<tmn> using Buckwalter Arabic Morphological Analyzer (BAMA) produced 21 different analyses. Further, it was estimated that the average number of ambiguities for a token in the majority of languages is 2.3, whereas in Modern Standard Arabic, it is 19.2 (Farghaly and Shaalan, 2009). When the same process for the same word "ثمن "<tmn> using "AlKhalil Morpho Sys" is executed, the analyser determines 40 different analyses, considering all possible diacritical marks.

There are some approaches designed to achieve this disambiguation. The best known ones are statistical/probabilistic approaches, rule-based methods, and hybrid systems that using a combination of both statistical and rule-based methods:

- Statistical/probabilistic methods: Almost all of these are based on Markov models where training consists of learning both lexical and contextual probabilities. This approach is based on a large manually annotated corpus from which we extract probabilities.
- Rule-based methods: They function using rules that have been defined by linguists. A rule-based method is composed of three tasks:
  1. Morphological analysis: This consists of segmenting a sequence of input words into morphemes with respect to the language grammar. This process is accomplished by morphological analysers;
  2. Lexicon research: Lexicons include words that cannot be analysed in the morphological task, such as some stop words, proper nouns, Arabized nouns, and misclassified words;
  3. Sentence structure (El Hadj et al., 2009): This is based on the relationship between untagged words and their adjacent words. The Arabic language has relationships between adjacent words. For example, prepositions and interjections are usually followed by nouns. The word position in the sentence is an effective indicator to identify nouns. Some words always followed by nouns construct a linguistic rule to identify them in the text such as "إن وأخواتها", "كان وأخواتها ", and some of these words are mainly used when recognizing proper nouns such as "السيد " and "الجامعة " 'Mr. University'.
- Hybrid automatic system: This involves combining different methods such as rule-based methods with statistical/probabilistic methods. This system is used to assign the best tag for each of the words of the input text.

2.3. Arabic PoS taggers

A significant part of the work has been undertaken in the area of Arabic PoS tagging (Al-Sughaiyer and Al-Kharashi, 2004; Sawalha and Atwell, 2010); other projects have been developed by companies (Xerox, Sakhr, RDI) as commercial software. In this section, we summarize some of the most relevant works on PoS tagging.

The stochastic PoS taggers provide the appropriate tags based on the most likely tag sequence in tagged corpora; many developed algorithms are employed (Altabba et al., 2010), such as the Viterbi algorithm (Viterbi, 1967) using a Hidden Markov Model (HMM).

The most relevant PoS taggers based on this approach (Diab et al., 2004) are based on Support Vector Machine (SVM), a supervised learning algorithm that uses LDC's PoS tagset, consisting of 24 tags. Another SVM-based, Yamcha, which uses Viterbi decoding, was developed by Habash and Rambow (2005). The approach of Maamouri and Cieri (2002)) is based on the automatic annotation output produced by Tim Buckwalter's morphological analyser; it achieved an accuracy of 96%. Banko and Moore (2004) presented an HMM tagger that exploits context on both sides of a word to be tagged. It is evaluated in both the unsupervised and supervised cases and achieved an accuracy of approximately 96%. Another PoS tagger, similar to the one integrated into the Stanford PoS Tagger, adopted a maximum entropy approach by enriching the information sources used for tagging. Its end result accuracy on the Penn Treebank achieved 96.86% overall, and 86.91% on previously unseen words (Toutanova and Manning, 2000). Another probabilistic tagger was adapted for Arabic (Zeroual and Lakhouaja, 2016a); it differs from other probabilistic taggers in the manner the transition probabilities are estimated, namely with a decision tree. The authors report that the obtained accuracy rates were 99.4%, 92.6%, and 81.9% for the Quranic-vowelled corpus "Al-Mus'haf" (Zeroual and Lakhouaja, 2016b), unvowelled "Al-Mus'haf" corpus, and the NEMLAR corpus (Attia et al., 2005), respectively.

The Qutuf (Altabba et al., 2010) tagger is based on a system that consists of two tagging phases: premature and overdue (usual tagging). The premature tagging occurs before the morphological analysis phase, whereas the usual tagging happens after, and requires rules from a linguistic expert or manually annotated corpus to statistically generate the rules. The Qutuf tagset is based on the Sawalha tagset with refinement and expansion. Brill's "transformation-based" or "rule-based" PoS tagger for Arabic (Freeman, 2001) uses a machine-learning approach based on the Brown cor-