Is Arabic Part of Speech Tagging Feasible Without Word Segmentation?

Emad Mohamed, Sandra Kübler
Indiana University
Department of Linguistics
Memorial Hall 322
Bloomington, IN 47405
USA
{emohamed,skuebler}@indiana.edu

Abstract

In this paper, we compare two novel methods for part of speech tagging of Arabic without the use of gold standard word segmentation but with the full POS tagset of the Penn Arabic Treebank. The first approach uses complex tags without any word segmentation, the second approach is segmention-based, using a machine learning segmenter. Surprisingly, word-based POS tagging yields the best results, with a word accuracy of 94.74%.

1 Introduction

Arabic is a morphologically rich language, in which a word carries not only inflections but also clitics, such as pronouns, conjunctions, and prepositions. This morphological complexity also has consequences for the part-of-speech (POS) annotation of Arabic: Since words can be complex, POS tags refer to segments rather than to whole words. Thus, the word wsyrfEwnhA (in Buckwalter transliteration; engl.: *and they will raise it*) is assigned the following POS tag: [CONJ+FUTURE_PARTICLE+IMPERFECT_VERB_PREFIX +IMPERFECT_VERB+IMPERFECT_VERB_SUFFIX_MASCULINE_PLURAL_3RD_PERSON+OBJECT_PRONOUN_FEMININE_SINGULAR] in the Penn Arabic Treebank (ATB) (Bies and Maamouri, 2003); the boundaries between segments are depicted by + signs. Automatic approaches to POS tagging either must assign such complex tags from a large tagset to complete words, or they must segment the word first and then assign POS tags to the segments. Previous approaches (Diab et al., 2004; Habash and Rambow, 2005; van den Bosch et al., 2007; AlGahtani et al., 2009) chose the segmentation approach but concentrated on POS tagging by using the segmentation provided by the ATB. Additionally, Diab et al. and Habash and Rambow used a reduced tagset. Diab et al. and Habash and Rambow used Support Vector Machines, the former with a standard windowing approach, the latter performing a full morphological analysis before POS tagging. Van den Bosch et al., whose approach is the most similar to ours, used memory-based learning with the full ATB tagset. They report a POS tagging accuracy of 91.5% (93.3% on known words, 66.4% on unknown words). However, they also evaluated on words as defined in the ATB, which differs from written Arabic in the treatment of affixes with syntactic functions (see section 2 for details). AlGahtani et al. used transformation-based learning combined with a morphological analysis for unknown words and words containing clitics. They reached a POS tagging accuracy of 96.9% on ATB1. Surprisingly, their results are lower for the experiment using the whole ATB (96.1%).

In this paper, we present two methods for Arabic POS tagging that do not require gold standard segmentation but can rather be used for naturally occurring Arabic. We investigate two different approaches: (1) Assigning complete POS tags to whole words, without any segmentation, and (2) a segmentation-based approach, for which we developed a machine learning based segmenter. In this approach, the words are first passed to the segmenter, then to the POS tagger. The first approach is surprisingly successful given the complex-

ity of the task, reaching an accuracy on the word level of 94.74%, as compared to 93.47% for the segmentation-based approach. Thus, the result for the whole word approach is very close to the result obtained by using gold standard segmentation (94.91%). However, a more detailed analysis shows that this good performance of the word-based approach is due to its performance on known words while the few unknown words are more often misclassified: we reach an accuracy of 96.61% on known words but only 74.64% on unknown words.

2 Data, Methods, and Evaluation

Like the previous approaches, we base our experiments on the ATB, specifically on the after-treebank POS files, for extracting our training and test sets. More specifically, we use two sections of the ATB (P1V3 and P3V1) since those two sets do not contain duplicate sentences. This data set contains approximately 500,000 words. In order to be as representative of real-world Arabic, we use the non-vocalized version of the treebank. Since previous approaches, to our knowledge, used different data sets, our results are not directly comparable.

For both segmentation and POS tagging, we modified the ATB representation of words in order to obtain the text, as it would occur in newscasts. The ATB treats inflectional affixes, including the definite article `Al`, as part of a word but splits off those affixes that serve a syntactic function into separate words. In order to obtain text as it occurs in newscasts, we re-attached all conjunctions, prepositions, pronouns, and any elements that constitute parts of the word as an orthographic unit (with the exception of punctuation) to the word. The word `ltxbrh` (engl.: *in order to tell him*), for example, is represented as three words in the ATB, `l`, `txbr`, and `h`, but is treated as one single unit in our experiment. Our second modification concerns the null element in Arabic verbs. Since Arabic is pro-drop, the ATB annotation includes a null element in place of the omitted subject plus the POS tag it would receive. Since this information is not available in naturally occurring text, we delete the null element and its tag. For example, {`i$otaraY+(null)` and its tag PV+PVSUFF_SUBJ: 3MS would occur as {`i$otaraY` with the tag PV in our representation (we additionally remove the short vowels).

We perform 5-fold cross validation and use the same data split for all three types of experiments: (1) POS tagging using gold standard segmentation taken from the ATB, (2) POS tagging using a segmenter, and (3) POS tagging whole words with complex POS tags. The first experiment serves as the upper bound and as a comparison to previous approaches. The second experiment uses an automatic segmenter as a pre-processing component to the POS tagger. This means that the accuracy of the segmenter is also the upper limit of the POS tagger since errors in segmentation inevitably lead to errors in POS tagging. The last experiment uses full words and complex POS tags. The purpose of this experiment is to determine whether it is possible to tag complete words without segmentation.

The segmenter and the two POS taggers use memory-based learning. For segmentation, we use TiMBL (Daelemans and van den Bosch, 2005); for POS tagging MBT, a memory-based tagger (Daelemans et al., 1996). Memory-based learning is a lazy learning paradigm that does not abstract over the training data. During classification, the k nearest neighbors to a new example are retrieved from the training data, and the class that was assigned to the majority of the neighbors is assigned to the new example. MBT uses TiMBL as classifier; it offers the possibility to use words from both sides of the focus word as well as previous tagging decisions and ambitags as features. An ambitag is a combination of all POS tags of the ambiguity class of the word.

Word segmentation is defined as a per-letter classification task: If a character in the word constitutes the end of a segment, its class is '+', otherwise '-'. We use a sliding window approach with 5 characters before and 5 characters after the focus character, the previous decisions of the classifier, and the POS tag of the focus word assigned by the whole word tagger (cf. below) as features. The best results were obtained for all experiments with the IB1 algorithm with similarity computed as weighted overlap, relevance weights computed with gain ratio, and the number of k nearest neighbors equal to 1.

For POS tagging, we use the full tagset, with information about every segment in the word, rather than the reduced tagset (RTS) used by Diab et al. and Habash and Rambow, since the RTS assumes

Gold Standard Segmentation		Segmentation-Based Tagging		Whole Words
SAR	WAR	SAR	WAR	WAR
96.72%	94.91%	94.70%	93.47%	94.74%

Table 1: POS tagging results.

a segmentation of words in which syntactically relevant affixes are split from the stem. The word w+y+bHv+wn+hA, for example, in RTS is split into 3 separate tokens, w, ybHvwn, hA. Then, each of these tokens is assigned one POS tag, Conjunction for w, Imperfective Verb for ybHvwn, and Pronoun for hA. The split into tokens makes a preprocessing step necessary, and it also affects evaluation since a word-based evaluation is based on one word, the RTS evaluation on 3 tokens for the above example.

For all the POS tagging experiments, we use MBT. The best results were obtained with the Modified Value Difference Metric as a distance metric and with k = 25. For known words, we use the IGTree algorithm and 2 words to the left, their POS tags, the focus word and its ambitag, 1 right context word and its ambitag as features. For unknown words, we use IB1 as algorithm and the unknown word itself, its first 5 and last 3 characters, 1 left context word and its POS tag, and 1 right context word and its ambitag tag as features.

3 Experimental Results and Discussion

3.1 Word Segmentation

The memory-based word segmentation performs very reliably with a word accuracy of 98.23%. This also means that when the segmentation module is used as a pre-processing step for POS tagging, the accuracy of the tagger will have this accuracy as its upper bound. While there are cases where wrong segmentation results in the same number of segments, all of these words were assigned the wrong POS tags in our data. In an error analysis, we found that words of specific POS are more difficult to segment than others. Proper nouns constitute 33.87% of all segmentation errors, possibly due to the fact that many of these are either foreign names that resemble Arabic words (e.g. Knt, which is ambiguous between the English name *Kent*, and the Arabic verb *I was*), or they are ordinary nouns used as proper nouns but with a different segmentation (e.g. AlHyAp, engl.: *the life*). The POS tag with the second highest error rate was the noun class with 30.67%.

3.2 Part of Speech Tagging

Table 1 shows the results of the three POS tagging experiments described above. For the segmentation-based experiments, we report per-segment (SAR) and per-word (WAR) accuracy. As expected, POS tagging using gold standard segments gives the best results: 94.91% WAR. These results are approximately 3 percent points higher than those reported by van den Bosch et al. (2007). Although the results are not absolutely comparable because of the different data sets, this experiment shows that our approach is competitive. The next experiments investigate the two possibilities to perform POS tagging on naturally occurring Arabic, i.e. when gold segmentation is not available. The results of these experiments show that POS tagging based on whole words gives higher results (WAR: 94.74%) than tagging based on automatic segmentation (WAR: 93.47%). This result is surprising given that tagging whole words is more difficult than assigning tags to segments, as there are 993 complex tags (22.70% of which occur only once in the training set), versus 139 segment tags. A detailed error analysis of a previous but similar experiment can be found in Mohamed and Kübler (2010).

We assume that these results are an artifact of the ATB since it is based exclusively on newswire texts. This means that there is only a limited vocabulary, as shown by the very low rate of unknown words: across the five folds, we calculated an average of 8.55% unknown words. In order to test our hypothesis that unknown words are tagged more reliably with a segment-based approach, we performed an analysis on known and unknown words separately. The results of this analysis are shown in Table 2.

This analysis shows that for all experiments, the unknown words are tagged with a considerably

	Gold Standard Segmentation	Segmentation-Based Tagging	Whole Words
Known words	95.90%	95.57%	96.61%
Unknown words	84.25%	71.06%	74.64%

Table 2: POS results for known and unknown words.

lower accuracy. However, the loss of performance is more pronounced in the approaches without gold segmentation. It is also evident that tagging whole words reaches a higher accuracy than segment-based tagging for both known words and unknown words. From these results, we can conclude that while segmentation makes properties of the words available, it is not required for POS tagging. We also investigated the poor performance of the segmentation-based tagger. A closer look at the results for unknown words in segmentation-based tagging shows that 59.68% of the tagging errors are direct results from incorrect segmentation decisions. In comparison, for known words, only 6.24% of the incorrectly tagged words are also ill-segmented. This means that even though the quality of the segmenter is very high, the errors still harm the POS tagging step.

To make our results more comparable to those by Habash and Rambow (2005), we converted the test set with the POS tags from the whole word tagger to their tokenization and to a reduced tagset of 15 tags. In this setting, we reach a tokenization accuracy of 99.36% and a POS tagging accuracy of 96.41%. This is very close to the results by Habash and Rambow so that we conclude that high accuracy POS tagging for Arabic is possible without a full morphological analysis.

4 Conclusions and Future Work

We have presented a method for POS tagging for Arabic that does not assume gold segmentation, which would be unrealistic for naturally occurring Arabic. The approach we developed is competitive although it uses the full POS tagset, without any previous morphological analysis. The results of our experiments suggest that segmentation is not required for POS tagging. On the contrary, using whole words as basis for POS tagging yields higher accuracy, thus rendering a full morphological analysis or segmentation unnecessary. We reached the best results in tagging whole words both for known words and unknown words. These results were only marginally worse that the results obtained by the experiment based on gold segmentation.

The weakness of the segmentation-based approach is its low accuracy on unknown words. In the future, we will investigate knowledge-richer methods for segmentation. In particular, we will investigate whether an automatic vocalization step previous to segmentation will improve POS tagging accuracy for unknown words.

References

Shahib AlGahtani, William Black, and John McNaught. 2009. Arabic part-of-speech-tagging using transformation-based learning. In *Proceeedings of the 2nd International Conference on Arabic Language Resources and Tools*, Cairo, Egypt.

Ann Bies and Mohamed Maamouri. 2003. Penn Arabic Treebank guidelines. Technical report, LDC, University of Pennsylvania.

Walter Daelemans and Antal van den Bosch. 2005. *Memory Based Language Processing*. Cambridge University Press.

Walter Daelemans, Jakub Zavrel, Peter Berck, and Steven Gillis. 1996. MBT: A memory-based part of speech tagger-generator. In Eva Ejerhed and Ido Dagan, editors, *Proceedings of the 4th Workshop on Very Large Corpora*, pages 14–27, Copenhagen, Denmark.

Mona Diab, Kadri Hacioglu, and Daniel Jurafsky. 2004. Automatic tagging of Arabic text: From raw text to base phrase chunks. In *Proceedings of HLT-NAACL*, Boston, MA.

Nizar Habash and Owen Rambow. 2005. Arabic tokenization, part-of-speech tagging and morphological disambiguation in one fell swoop. In *Proceedings of ACL-2005*, pages 573—580, Ann Arbor, MI.

Emad Mohamed and Sandra Kübler. 2010. Arabic part of speech tagging. In *Proceedings of LREC*, Valetta, Malta.

Antal van den Bosch, Erwin Marsi, and Abdelhadi Soudi. 2007. Memory-based morphological analysis and part-of-speech tagging of Arabic. In Abdelhadi Soudi, Antal van den Bosch, and Günter Neumann, editors, *Arabic Computational Morphology*. Springer.