Arabic Part of Speech Tagging

Emad Mohamed, Sandra Kübler

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

Abstract

Arabic is a morphologically rich language, which presents a challenge for part of speech tagging. In this paper, we compare two novel methods for POS 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 that describe full words and does not require any word segmentation. The second approach is segmentation-based, using a machine learning segmenter. In this approach, the words are first segmented, then the segments are annotated with POS tags. Because of the word-based approach, we evaluate full word accuracy rather than segment accuracy. Word-based POS tagging yields better results than segment-based tagging (93.93% vs. 93.41%). Word based tagging also gives the best results on known words, the segmentation-based approach gives better results on unknown words. Combining both methods results in a word accuracy of 94.37%, which is very close to the result obtained by using gold standard segmentation (94.91%).

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: Because of the morphological complexity, a single stem may correspond to thousands of different word forms, which leads to data sparseness issues. Additionally, since words can be complex, POS tags contain information about the morphology, i.e. they refer to segments rather than to whole words. Thus, the word وسيكتبونها (wsyktbwnhA in Buckwalter transliteration, engl.: *And they will write 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]

This word form consists of a conjunction, a future particle, an inflectional prefix the verb stem, an inflectional suffix, and a pronominal object. The boundaries between segments are depicted by + signs. As can be seen from this example, three of the segments (the conjunction, the future particle and the object pronoun) as well as the stem ktb, are syntacticlly independent although they are part of the orthographic form, i.e. they are clitics.

Automatic approaches to POS tagging either mus 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 Penn Arabic Treebank (ATB) (Bies and Maamouri, 2003).
In this paper, we present two methods for Arabic POS tagging that do not require gold standard data:

- Whole word tagging: In this approach, we assign complete POS tags to whole words, without any segmentation.
- Segmentation-based tagging: For this approach, we developed a machine learning based segmenter. Thus, the words are first passed through the segmenter, then the segmented text is passed to the POS tagger, and each segment is assigned its POS tag.

The first approach is surprisingly successful given the complexity of the task, reaching 93.93%, as compared to 93.41% for the segmentation-based approach. 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. When we combine the two approaches, i.e. use full words when they occur in the training data and segments otherwise, we obtain an accuracy of 94.37% on all words, which is very close to the result obtained by using gold standard segmentation (94.91%).
In the remainder of the paper, we first give an overview of characteristics of the Arabic language that are relevant for our investigation (cf. section 2.). Then, we describe previous approaches to POS tagging Arabic in section 3. In section 4., we discuss the data set that we use and the experimental setup, and in section 5., we discuss our results.

2. The Arabic Language

Like in other Semitic languages, Arabic words, here defined as whitespace delimited units, are complex units made up of a stem plus inflections (to which we refer as the base form). Attached to the base form are usually clitics that can be either proclitics of the set `w,f` which denote conjunctions, the set `k,l,b`, which are prepositions, and the set `s,l` which are verbal proclitics meaning *shall/will* and *in order to* respectively. Enclitics can either be object pronouns, possessive pronouns, or personal pronouns. Each form is ambigous between those and can be distinguished by the nature of the host. For example, `hA` is the feminine 3rd person clitic that is an object pronoun when its host is a verb or a verbal noun e.g. `rAfqhA` (engl.: *He escorted her*), and a possessive pronoun when the host is a noun, e.g. `syArthA` (engl.; *Her car*). Nominal inflections, which also occur with adjectives, denote person, gender, case, and definiteness, while verbal inflections mark tense, mood, aspect, number, gender and voice.

For example, the word `wsywqEAnhA` has four tokens, i.e. parts with syntactic functions: `w` (engl.: *and*) `s` (engl.: *will*) `ywqEAn` (engl.: *They both sign*), and `hA` (engl.: *it*). The base form `ywqEAn` contains the stem `wqE` (engl.: *sign*) as well as the inflections `y` (present tense/masculine/3rd person) and `An` (dual/indicative).

In this paper, we present both segmentation, which includes determining the boundaries between all the word parts, i.e. inflections, stems, or clitics, and tokenization, which determines the boundaries between syntactically functional units to the exclusion of inflectional affixes. The word above would then be segmented as `w+s+y+wqE+An+hA`, but tokenized as `w+s+ywqEAn+hA`.

Whether we choose to perform segmentation, tokenization, or use whole words for POS tagging will affect the choice of the POS tagset. Tagging segments, one the one hand, requires that we provide tags for all inflectional segments, which will result in a medium sized tagset (139 segments tags), but the segments themselves will be rather ambiguous. Tagging tokens, on the other hand, may not require such fine-grained tags; and the tagsets that have been mostly used so far, which are token tagsets, have been smaller in number (15 – 25). Tagging words, in contrast, results in often unambiguous words but in the largest tagset of 993 complex tags (see section 1. for an example of such a complex tag).

3. Previous Work

As mentioned above, previous approaches (Diab et al., 2004; Habash and Rambow, 2005; van den Bosch et al., 2007; AlGahtani et al., 2009) use the segmentation provided by the Penn Arabic Treebank for POS tagging.

Diab et al. (2004) use a machine learning approach, Support Vector Machines, to model Arabic Part of Speech Tagging as a classification approach using a manually reduced tagset, which maps Arabic to 24 POS tags similar to those used in the Penn Treebank (Santorini, 1990) for English. Their feature set includes the focus word in a window of two words to the right and left, character n-grams of the focus word, the types of the words in terms of alphabetic and numeric characters, and previous tagging decisions for the words within the left context. Diab et al. report an accuracy of 95.5% on all tokens drawn from the ATB, part 1 version 1 (ATB1).

Habash and Rambow (2005) follow Diab et al. in using SVM's for Arabic POS Tagging, but they use a full morphological analyzer, instead of plain classification, to produce all the possible morphological forms of a certain word. In a second step, the classifier decides between all analyses produced by the morphological analyzer. This means that POS tagging is a by-product of the morphological analysis. Habash and Rambow criticize the tagset used by Diab et al. as unmotivated, since it makes distinctions based on English that may not be relevant for Arabic; Habash and Rambow use this tagset for comparison reasons along with a smaller tagset of 15 POS tags. They report that the use of the morphological analyzer improves POS tagging accuracy; their POS evaluation shows an accuracy of 97.6% on ATB1 and an accuracy of 95.7% on ATB2, both based on gold standard tokenization.

Van den Bosch et al. (2007) use memory-based learning (Aha et al., 1991) for both morphological analysis and POS tagging of Arabic. For POS tagging, they use MBT, a memory-based tagger (Daelemans et al., 1996). Unlike Habash and Rambow and Diab et al., they use whole words in their approach, i.e. they use the segmentation as given in the ATB, which means that conjunctional prefixes and pronominal suffixes appear as separate words, which is not the case for naturally occurring Arabic. Van den Bosch et al. use ATB1 in an 11-fold cross validation. They report an overall accuracy of 91.5% with a 93.3% accuracy on known words and 66.4% accuracy on unknown words. This approach is the closest to ours, which also uses MBT, but a different data set.

AlGahtani et al. (2009) use transformation-based learning as implemented in the Brill tagger (Brill, 1994) for POS tagging Arabic with segment-based tags. For training, they use the gold standard segmentation of the ATB, while in testing, segmentation is performed by the Buckwalter morphological analyzer (Buckwalter, 2004). AlGahtani et al. use bigram information from the morphological analyses to select the preferred one, which is then passed to the Brill tagger. AlGahtani et al. (2009) evaluate their approach on the whole ATB as well as on ATB1. For ATB1, they reach a POS tagging accuracy of 96.9%, which is between the results of Diab et al. and Habash and Rambow. However, it is surprising that their results are lower for the experiment using the whole ATB (96.1%), even though large parts of the treebank are duplicated between parts, so that it is likely that parts of their test set were actually present in the training set.

4. Data, Methods, and Evaluation

4.1. Data Set

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 used 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 the previous approaches 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. For this reason, all conjunctions, prepositions, pronouns, and any elements that constitute parts of the word as an orthographic unit (with the exception of punctuation) are re-attached to the word. The word لتخبره (`ltxbrh`, engl.: *in order to tell him*), for example, is represented as three entries in the ATB, `l`, `txbr`, and `h`, but is treated as one single unit in our experiment. Another 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 (but we additionally remove vocalization).

4.2. Experimental Setup

We perform a 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 of whole words with complex POS tags. The first experiment serves as the upper bound for the segmentation-based approach and as a comparison to previous approaches. The second experiment uses an automatic segmenter as a preprocessing 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.

Both the segmenter and the two POS taggers are based on memory-based learning. For the segmenter, we use TiMBL (Daelemans and van den Bosch, 2005; Daelemans et al., 2007); 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 in the background; 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.

While Diab et al. (2004) performed tokenization, we concentrate on segmentation. This means that we determine the boundary of every segment, whether inflectional or clitic. Diab et al., in contrast, only split off tokens that have syntactic functions, such as conjunctions, prepositions, and pronouns. For example, the sentence في خدمة مآربهم ومصالحهم الضيقة بضيق أفقهم is segmented in the following way: `fy xdm+p m|rb+hm w+mSAlH+hm Al+Dyq+p b+Dyq >fq+hm`. The tokenization by Diab et al., in contrast, renders this sentence as `fy xdmp m|rb hm w mSAlH hm AlDyqp b Dyq >fq hm`. I.e. the sentence of 7 words consists of 15 segments or 12 tokens.

For our purpose, we define word segmentation as a per-letter classification task in which segment boundaries are marked with a '+': 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 as features. The best results were obtained for all experiments with the IB1 algorithm with similarity computed as overlap, using weights based on gain ratio, and the number of k nearest neighbors equal to 1. These settings were determined in a non-exhaustive search.

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 we are convinced that the information omitted by the RTS is important for tasks building on POS tagging. The word `y+bHv+wn`, for example, is assigned the RTS tag of VBP (Imperfective Verb), neglecting the masculine plural specification.

For all the POS tagging experiments, we use MBT. We optimized the feature and parameter settings in a non-exhaustive search. The best results were obtained with the Modified Value Difference Metric (MVDM) as a distance metric and with k = 25. For known words, we use the IGTree algorithm and two words to the left, their POS tags, the focus word and its ambitag, one right context word and its ambitag as features. For unknown words, we use IB1 as algorithm and the unknown word itself, its first and last three characters, one left context word and its POS tag, and one right context word and its POS tag as features.

4.3. Evaluation

Segmentation is evaluated by calculating accuracy on whole words. This means, a word is counted as correct if and only if all the segment boundaries were assigned correctly, according to the gold standard segmentation even if the segmentation provided by the classifier is plausible. For example, let us assume that the word `w+Al+ml*+At` is segmented as `w+Al+ml+*At`. Although two segments are

Baseline	Gold Standard Segmentation		Segmentation-Based Tagging		Whole Words
WAR	SAR	WAR	SAR	WAR	WAR
77.02%	96.72%	94.91%	94.60%	93.41%	93.93%

Table 1: POS tagging results.

correctly identified, the whole word is counted as wrong. Also, the word txft can either be segmented as t+xft or txf+t, and is thus ambiguous with regard to segmentation, but it it will be counted as correct only if the segmentation provided by the segmenter matches the segmentation of the word in context as given in the gold standard.

Previous POS tagging experiments were based on tokens (cf. section 3.). As a consequence, POS tagging accuracy was reported on those tokens as well. Since we compare a full word tagging approach with one based on segments, a purely segment-based accuracy is not feasible. For this reason, we report word accuracy rates (WAR) for all experiments. Word accuracy is defined as the number of words that are tagged correctly. This means that for the segment-based tagging, we reattach the segments and their POS tags before evaluation. Where applicable, we also report segment accuracy rates (SAR).

To our knowledge, we are the first to report word accuracy rates. However, we believe that this measure of accuracy, while giving lower accuracy rates, provides a more realistic picture of POS tagging accuracy.

5. Experimental Results and Discussion

5.1. Word Segmentation

The memory-based word segmentation performs very reliably with a word accuracy of 98.15%. While segmentation quality is high, there is a significant difference between known words and unknown words. On the one hand, known word segmentation accuracy averages 99.74% across 5 folds. Unknown words, on the other hand, reach a considerably lower average accuracy of 81.39%. This indicates that the quality of segmentation, as well as any process that depends on it, such as POS tagging, will depend on the percentage of unknown words in the text. 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 31.82% 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 form *I was*), or they are ordinary nouns used as proper nouns but with a different segmentation(e.g. AlhyAp). The next most frequent category are nouns with 3.60%.

5.2. Part of Speech Tagging

Table 1 shows the results of the three POS tagging experiments described above as well as for a baseline. For the baseline, we use the most basic experiment: POS-tagging using whole words and choosing the most frequent tag for each word in the training set. Unknown words are assigned the most frequent tag in the training set, NOUN. This yields an accuracy of 77.02%. 8.5% of all the words in the test set were unseen, i.e. they did not occur in the training set. The baseline accuracy is rather high because most words are unambiguous, and the average number of POS tags per word is 1.06. However, the accuracy is considerably lower than the baseline reported by Diab et al. (2004), who report an accuracy of 92.2% for a similar baseline. However, their experiments are based on tokens, not on complete words, and they use a reduced tagset of 24 POS tags as compared to the 993 complex tags of the full tagset that we use in our experiments.

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), which may be due to the larger training set. 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: 93.93%) than tagging based on automatic segmentation (WAR: 93.41%). 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), versus 139 segment tags.

Figure 1 shows an example of a sentence with the gold standard POS tags as well as the tags assigned by the whole word tagger and the segmentation-based tagger. The sentence contains two out-of-vocabulary words: m|rbhm and >fqhm. The whole word approach tags both correctly (without any use of word segmentation) as well as it does the known words. The segmentation-based approach fails to tag the latter word correctly due to a segmentation error. The correct segmentation of >fqhm is >fq+hm, which is tagged as a noun plus a possessive pronoun for the 3rd person masculine plural. But it is segmented as >+fqhm and consequently tagged as a verb prefix plus an imperfect verb. While segmentation-based tagging has been shown to perform better on unknown words, this is not always the case as witnessed by this example.

One explanation for these results is that segments are more ambiguous than whole words. The following shows an example where the whole word is unambiguous while the segments are ambiguous, and the POS tagger made wrong decisions. The often cited word wbhsnAthm (engl.:

Word	Gold POS tag	Word-based POS tag	Segmentation-based POS tag
fy	PREP	PREP	PREP
xdm+p	NN+NSUFF_FEM_SG	NN+NSUFF_FEM_SG	NN+NSUFF_FEM_SG
m—rb+hm	NN+POSS_PRON_3MP	NN+POSS_PRON_3MP	NN+POSS_PRON_3MP
w+mSAlH+hm	CONJ+NN+POSS_PRON_3MP	CONJ+NN+POSS_PRON_3MP	CONJ+NN+POSS_PRON_3MP
Al+Dyq+p	DET+ADJ+NSUFF_FEM_SG	DET+ADJ+NSUFF_FEM_SG	DET+ADJ+NSUFF_FEM_SG
b+Dyq	PREP+NN	PREP+NN	PREP+NN
>fq+hm	NN+POSS_PRON_3MP	NN+POSS_PRON_3MP	IV1S+IV

Figure 1: Example of a sentence with gold standard POS tags and the tags assigned by the taggers. (For reasons of readability, we have shortened that POS tag NOUN to NN.)

segment	Possible POS Tags
w	CONJUNCTION, PREPOSITION, ABBREV
b	PREP, ABBREV
Hsn	NOUN, ADJ, NOUN_PROP, PV
At	NOUN_SUFFIX_FEMININE_PLURAL, ADJ
hm	NOUN, POSS_PRON_3MP, IV, IVSUFF_DO:3MP, PRON_3MP, PVSUFF_DO:3MP

Table 2: POS tags for individual segments.

and+by+their+virtue+s) is not ambiguous by and of itself, the only possible tag being CONJ+PREP+NOUN+ NOUN_SUFFIX_FEMININE_PLURAL+POSS_PRON_ 3MP. However, the individual segments are highly ambiguous with 3 x 2 x 4 x 2 x 6 = 288 possible composite tags for the word (see Table 2). This means that POS tagging for Arabic needs to walk a fine line between data sparseness for full words, and high ambiguity for segments.

As a consequence of this observation, 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 8.5% 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 3.

This analysis shows that for all experiments, the unknown words are tagged with a considerably lower accuracy. However, the loss of performance is more pronounced in the approaches that do not rely on gold segmentation. It is also evident that while tagging whole words reaches a higher accuracy than segment-based tagging for known words, unknown words are tagged more reliably by the segment-based approach. We can therefore conclude that segment-based POS tagging is more suitable for texts with a higher percentage of unknown words. A closer look at the results for unknown words in segmentation-based tagging shows that 59.68% of the tagging errors are results from incorrect segmentation decisions. In comparison, for known words, only 6.24% of the incorrectly tagged words are also ill-segmented.

The results in Table 3 also allow the conclusion that we can improve results further by combining the best of the two approaches, which is corroborated by the following experiment: If we use whole words for known word tagging, and segmentation-based tagging for unknown words, we can reach a word accuracy of 94.37%, which is higher than both results for the individual experiments and very close to the results obtained by using gold-standard segments.

5.3. Error Analysis

We performed an error analysis on the whole word tagged data set in order to understand better where the challenges lie. The most common confusion in our data is that between nouns (NOUN) and adjectives (ADJ). Nouns are assigned the ADJ tag in 7.88% of all errors and adjectives are assigned the NOUN tag 7.75% of all errors. This confusion set alone is thus responsible for over 15% of all errors. The noun-adjective distinction is not all that clear in Arabic since adjectives can be used as nouns, and they behave morphologically in the same way as nouns do: Adjectives receive the plural and feminine markers as well as case in the same way that nouns do. This corresponds to the observation by (Diab et al., 2004) that the two categories cannot be separated cleanly in Arabic, which leads to inconsistencies in the treebank annotation.

The second most common confusion in our data is that between proper nouns (NOUN_PROP) and nouns (NOUN). Proper nouns are tagged as NOUN in 9.1% of all errors. The reason behind this may be that proper nouns have the same distribution as nouns as they occur in the same texts. Arabic does not designate proper nouns in a specific way, for example, by capitalization as in English. Also, the determiner proclitic `Al` is attached to nouns in general, whether proper or not. Nouns tagged as NOUN_PROP makes up 2.51% of all errors.

The most interesting error type for our investigation are POS errors resulting from segmentation errors. Such errors constitute 28% of all errors, and result when a word is assigned more or fewer segments than it has. For example, the word `knt`, a proper noun, was segmented as `kn+t`, and was thus given two POS segment tags instead of one. In

	Baseline	Gold Standard Segmentation	Segmentation-Based Tagging	Whole Words
Known words	85.57%	95.90%	95.55%	96.60%
Unknown words	3.67%	84.25%	70.50%	65.50%

Table 3: POS results for known and unknown words.

	H&R ATB1	H&R ATB2	whole word tagger
Tokenization accuracy	99.1	–	99.33
POS accuracy	98.1	96.5	96.41

Table 4: Tokenization and POS results based on the reduced tagset of Habash and Rambow (2005).

our data, there are 2 201 words with extra segment errors and 1 089 words with fewer segment errors. This shows that the segmenter has a tendency to propose too many segments, which is rather surprising given that the training set contains a majority of examples where not to segment.

5.4. Comparison with the Results by Habash and Rambow

As mentioned before, our results are not directly comparable with previous results since we use the full tagset and have to evaluate on whole words. In order to make our results more comparable to the results by Habash and Rambow (2005), we converted the test set with the POS tags assigned by the whole word POS tagger and converted our segmentation to their tokenization and the full set of POS tags to a reduced tagset of 15 tags. All decisions were made so that the resulting data resembled the one by Habash and Rambow as closely as possible. For most cases, the conversion was straightforward, involving the deletion of morphological information and the combination of related POS tags. In a small number of cases, we made decisions from which we assume that they follow their procedure. One of the cases where we had to make such a decision involved the POS tag for possessive pronoun (POSS). Since the reduced tagset only has only a PRO tag for nominal pronoun, we decided to convert all POSS tags into PROs.

While this conversion gives us data in a very similar format to that by Habash and Rambow (2005), note that the comparison is based on different training and test sets. The results of this evaluation are shown in Table 4. The first row gives the accuracy of tokenization, the second row gives the POS accuracy, both evaluated on the word level. The results show that when we convert our segments to tokens (i.e. only parts that have their own syntactic function), our tokenization accuracy is slightly higher than Habash and Rambow's. The accuracy of our whole word tagger is very close to their results, even though it does not reach them completely. However, these results show that a high quality in POS tagging for Arabic can be reached without morphological analysis, even if it may require more training data.

6. Conclusion and Future Work

We have presented a method for POS tagging of 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. However, a direct comparison to previous work is difficult since there is no standard for splitting the data set into training and test data. The results of our experiments suggest that when the number of unknown words is large, performing automatic segmentation is very useful. In contrast, when there is a limited number of unknown words, 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 by combining whole word tagging for known words and segmentation-based tagging for unknown words, which yields results very close to the ones obtained by the experiment based on gold segmentation.

One of the weaknesses of the segmentation-based approach is its low accuracy on unknown words when compared to gold standard segmentation. 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 the accuracy for unknown words.

7. References

David Aha, Dennis Kibler, and Marc K. Albert. 1991. Instance-based learning algorithms. *Machine Learning*, 6:37–66.

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.

Eric Brill. 1994. A report of recent progress in transformation-based error-driven learning. In *Proceedings of the AAAI 94*, pages 722–727.

Tim Buckwalter. 2004. Arabic morphological analyzer version 2.0. Linguistic Data Consortium.

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.

Walter Daelemans, Jakub Zavrel, Ko van der Sloot, and Antal van den Bosch. 2007. TiMBL: Tilburg memory based learner – version 6.1 – reference guide. Technical Report ILK 07-07, Induction of Linguistic Knowledge, Computational Linguistics, Tilburg University.

Mona Diab, Kadri Hacioglu, and Daniel Jurafsky. 2004. Automatic tagging of Arabic text: From raw text to base phrase chunks. In *Proceedings of the 5th Meeting of the North American Chapter of the Association for Computational Linguistics and Human Language Technologies Conference, 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 the 43rd Annual Meeting of the ACL*, pages 573—580, Ann Arbor, MI.

Beatrice Santorini. 1990. Part-of-speech tagging guidelines for the Penn Treebank Project. Department of Computer and Information Science, University of Pennsylvania, 3rd Revision, 2nd Printing.

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.