Morphological Segmentation and Part of Speech Tagging for Religious Arabic

Emad Mohamed
Carnegie Mellon University Qatar
emohamed@qatar.cmu.edu

Abstract

We annotate a small corpus of religious Arabic with morphological segmentation boundaries and fine-grained segment-based part of speech tags. Experiments on both segmentation and POS tagging show that the religious corpus-trained segmenter and POS tagger outperform the Arabic Treebak-trained ones although the latter is 21 times as big , which shows the need for building religious Arabic linguistic resources. The small corpus we annotate improves segmentation accuracy by 5% absolute (from 90.84% to 95.70%), and POS tagging by 9% absolute (from 82.22% to 91.26) when using gold standard segmentation, and by 9.6% absolute (from 78.62% to 88.22) when using automatic segmentation.

1 Introduction

Traditional religious Arabic is the language variety used in pre-Modern texts dealing with the Quran, prophetic traditions, and the various books on Islamic law, Quran interpretation, Islamic philosophy and many other fields. It has more or less the same structure as Modern Standard Arabic but contains lexical items and some grammatical structures that may be out of place in today's newswire language. This has the potential of being incompatible with the NLP resources developed for Modern Standard Arabic, which are usually trained on newswire text.

In this paper, we annotate a small corpus of religious Arabic covering three religious domains, with fine-grained morphological segmentation boundaries and segment-based Part of Speech Tagging.

We show that even though the religious corpus is 21 times smaller than the Arabic Treebank sections used in this paper, the segmenter and POS tagger developed using the religious corpus yield much better results than those trained on the ATB. Moreover, a training set that is the concatenation of both the ATB and the religious corpus yields only slightly better results, which shows the need for building a religious Arabic Treebank. Small as it is, the religious corpus we annotate improves segmentation accuracy by 5% absolute (from 90.84% to 95.70%), and POS tagging by 9% absolute (from 82.22% to 91.26) when using gold standard segmentation, and by 9.6% absolute (from 78.62% to 88.22) when using automatic segmentation.

The rest of this paper is divided as follows: Section 2 presents the data we annotated and used in this paper, the methods and the evaluation schemes, section 3 presents the experiments we ran to test the usefulness of the religious corpus, and section 4 concludes and suggests future directions.

2 Data, Methods, and Evaluation

The author of this paper has annotated 3 booklets that cover religious material of enough variety to achieve proper coverage given the small amount of data included. The language variety these texts is written in is more of Classical Arabic than Modern Standard Arabic, and hence the need for the data. The books comprising the data are as follows: (1) ***Al>HAdyv Alnwwyp*** (الأحاديث النووية). This is a book of 50 traditions by Prophet Mohamed selected by Imam Nawawy (1233-1277 AC). The traditions cover a variety of topics with sayings attributed to Prophet Mohamed (571-631 AC). The book will henceforth be referred to as **Nawawy**.

(2) ***mtn >by $jAE*** (متن أبي شجاع) , **Matn** henceforth, is a booklet by the scholar ***>by $jAE*** (-1196 AC) about Islamic law that was intended to be short enough to be memorized by students. The book covers everything from cleanliness to Jihad, and from prayers to adjudication. The book is written in a very concise language.

(3) ***Almnq* mn AlDlAl*** (المنقذ من الضلال), (Eng. The Deliverer from Error), **Munqith** henceforth, is a book by Imam Gazaly (1058-1111) in which he narrates his journey to Sophism. The book focuses on matters of philosophy and belief. It is written in the first person, and addresses virtual listeners.

Table 1 provides basic statistics about the three books.

Book	**Words**	**Types**	**Segments**	**seg types**
Nawawy	4479	1323	6785	951
Matn	8832	3525	16774	2205
Munqith	14131	4824	23495	2857
Total	27442	**8686**	47054	**4818**

Table 1: basic statistics about the religious corpus

The three books above have been semi-automatically morphologically segmented and post-tagged by the author of this paper. First, the texts were automatically segmented and tagged then manually checked and corrected. The annotation scheme follows that of the Arabic Treebank (Bies and Mamouri, 2003). The annotation was meant to be as detailed as possible since detailed annotation can be used for deriving many forms of POS tags and word segmentations. The following section details both segmentation annotation and POS annotation.

2.1. Segmentation Annotation

For segmentation annotation, every possible affix, whether inflectional or clitical has been marked as a segment boundary. For example, the word ***fhjrth*** (فهجرته) is annotated as ***f+hjr+t+h***, where ***f*** is a syntactic token, ***hjr*** is a lexical unit, ***t*** is a subject inflection, and the final ***h*** is a pronoun. If the segmentation is ambiguous, then it is done according to the context.

2.2. Part of Speech Annotation

In annotating POS tags, we have opted for a tag set that is as detailed as possible. The tag set works at the segment level and encodes NUMBER, GENDER, DEFINITENESS, MOOD, CASE, and others. For example, the word ***fhjrth*** above is tagged as

f+/CONJ
hjr/NOUN
+t+/NSUFF_FEM_SG
+h/POSS_PRON_3MS

where ***CONJ*** means conjunction, ***NSUFF_FEM_SEG*** is the Noun Suffix for the Feminine Singular, and ***POSS_PRON_3MS*** is the Possessive Pronoun for the Third Person Masculine Singular. This process is highly context-dependent since the word ***fhjrth*** has at least four other possible POS tag sequences: f/**CONJ**+hjr/**PV**+t/**PVSUFF_SUBJ:3FS**+h/**PVSUFF_DO:3MS** , f/**CONJ**+hjr/**PV**+t/**PVSUFF_SUBJ:1S**+h/**PVSUFF_DO:3MS** , f/**CONJ**+hjr/**PV**+t/**PVSUFF_SUBJ:2MS**+h/**PVSUFF_DO:3MS** and /**CONJ**+hjr/**PV**+t/**PVSUFF_SUBJ:2FS**+h/**PVSUFF_DO:3MS**. This results from the fact that ***hjr*** is both a verb and a noun, ***t*** could be a first person subject pronoun, a second person female subject pronoun, a second person male subject pronoun, or, when affixed to a noun, a singular feminine marker.

2.3. Annotating Assimilated Forms

Arabic has some short (assimilated) forms consisting of a preposition and a pronoun. Table 2 list some of the most common forms in their long and short (naturally occurring) forms.

Our policy of annotating assimilated forms is to go with the conventional written form rather than undo the assimilation. For example, ***Emn*** is annotated as ***E***/PREP+***mn***/REL_PRONOUN instead of ***En***/PREP+***mn***/REL_PRON which is used in the ATB.

Long	**Short**	**English**
En mn	Emn	About whom
mn mn	Mmn	From whom
En mA	EmA	About what
mn mA	mmA	From What
ElY y	Ely	On me

Our policy of annotating assimilated forms is to go with the conventional written form rather than undo the assimilation. For example, ***Emn*** is annotated as ***E***/PREP+***mn***/REL_PRONOUN instead of ***En***/PREP+***mn***/REL_PRON which is used in the ATB.

A similar pattern occurs with the definite article ***Al*** when preceded by the preposition ***l***. While the ATB annotates this as ***l***/PREP+***Al***/DET as in the word ***l***/PREP+***Al***/DET+***mjtmE***/NOUN. We do not undo the assimilation and annotate this as ***l***/PREP+***l***/DET+***mjtmE***/NOUN as it occurs in naturally occurring Arabic.

The reason for this is that we do not make use of a morphological analyzer, and once they are segmented and tagged correctly, it's trivial to obtain the original information, although this is hardly needed.

The Arabic Treebank training set has been modified to conform to the same rules of assimilated forms.

3. Experiments

In order to show the usefulness of annotating religious data, we run the following three sets of experiments in which we vary the training set in both segmentation and POS tagging:

1. Train on newswire data and test on the religious data
2. Train on religious data and test
3. Train on a concatenation of the training sets in 1 and 2 above.

We divide the religious data into a training set (80% of the sentences) and a test set (20%). The sentences are assigned randomly to the test and training sets once, and then kept separate. This insures that the test set is the same across all experiments, which allows for proper comparisons between the different experiments.

3.1. Segmentation Experiments

For segmentation, we use the Timbl Memory-based learner (Daelemans *et al.*, 2010) with settings that have been tuned on the ATB data, with a feature representation in which we use the preceding five characters and the following five characters, when present, in a sliding window as features. We use the Timbl IB1 algorithm with similarity computed as overlap, using weights based on gain ratio, and the number of *k* nearest neighbours equal to 1. These settings were reported to achieve an accuracy of 98.15% when trained and tested on standard Arabic Treebank Data (Mohamed, 2010). These experiments also showed that the wider context and part-of-speech tags have only a very limited effect on segmentation quality and that word-internal context alone is enough for producing high quality segmentation.

We run three segmentation experiments:

1. **ATB**: In this experiment, we train on two sections of the ATB (p1v3+p3v2) and test on the religious test set.
2. **Religious**: we train on the Religious 80% and test on the religious 20%
3. **ATB+Religious**: We train on the concatenation of the training sets in 1 and 2, and test on the test set.

For evaluation, we use word level accuracy: a word is correctly segmented if and only if every segment boundary in it is marked correctly. A partially correct segmentation is a wrong segmentation. For example, the word ***f****hjrth* above has to receive the segmentation ***f****+hjr+t+h* to be considered correct, and even though ***fhjr+t+h*** has two segments marked correctly, the fact that one segment is wrong renders the whole word wrong.

3.1.1. Segmentation Results and Discussion

Table 3 shows the results of the three segmentation experiments above.

Experiment	**Accuracy**	**Known Word %**
ATB	90.84%	55.61
Religious	95.17%	76.70%
ATB+Religious	95.70	80.89%

Table 3: Segmentation Results

With the newswire data as training, the segmentation accuracy is 90.84%. A direct comparison with the Religious-trained segmenter shows a considerable difference of 4.33% in word accuracy. Combining both training sets (ATB+Religious) yields only a slight improvement of 0.53%.

There is a strong indication that the improvement may be attributed to the decrease in the rate of out-of-vocabulary words. While OOV's are 44.80% in the ATB experiment, they drop to 23.3 in the Religious experiment.

3.2. Part of Speech Tagging Experiments

For the POS tagging experiments, we use a memory-based tagger, MBT (Daelemans et al., 1996). The best results were obtained on the ATB data with the Modified Value Difference Metric as a distance metric and with ***k***, the number of nearest instances, = 25. For known words, we use the IGTree algorithm and 2 words to the left, their POS tags, the focus word and its ambitag (list of all possible tags), 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.

For POS tagging, we use two types of tagging settings:

1. Segmentation-based POS tagging in which the tagging is performed at the segment level. The words are then collected from those segments and the evaluation is performed at the word level. For example, to pos-tag the word ***llmmslmAt***, the word is first segmented into ***l+l+mslm+At***, and each segment is tagged (as in Table 4). Also note that While the segmentation used in the example in Table 4 is gold standard, we do not assume gold standard segmentation and will report results on both gold standard and automatic segmentations.

2. Whole Word Tagging. In this scheme, we do not use any segmentation but rather tag the word as a whole with a composite tag. The word *llmslmAt* thus receives the composite tag *PREP+DET+NOUN+NSUFF_FEM_PL* which has to be produced completely correctly by the tagger for the word to be correctly tagged.

Segment	Gold Tag	Predicted Tag
l	PREP	PREP
l	DET	DET
mslm	NOUN	ADJ
At	NSUFF_FEM_PL	NSUFF_FEM_PL
#	WORD_BOUNDARY	WORD_BOUNDARY
	l/PREP+l/DET+mslm/**NOUN**+At/NSUFF_FEM_PL	l/PREP+l/DET+mslm/**ADJ**+At/NSUFF_FEM_PL

Table 4: Segment-based tagging

The number of segment tags in the ATB training set is 139, while the number of tags in the Religious training set is 117. There are 6 tags in the Religious training set that do not occur in the ATB training set three of which are suffixes of the imperative verb. This shows the more conversational, albeit formal, nature of religious texts.

As far as the test set is concerned, it has 96 segment tags only one of them does not occur in the ATB training set, while 3 tags in the Religious training set do not occur in the test set.

Based on whether the training set comprises the ATB data alone, the religious training alone, or a combination thereof, we have run the following 9 experiments, six of which using segments and the other three with whole words:

1. **ATB GOLD**: Train on the ATB. The test segmentation is gold standard.
2. **ATB AUTO**: Train on the ATB. The test segmentation is automatic.
3. **REL GOLD**: Train on the Religious. The test segmentation is gold standard.
4. **REL AUTO**: Train on the Religious. The test segmentation is automatic.
5. **REL+ATB GOLD**: train on the concatenation of Religious and ATB, test on the gold standard segmentation
6. **REL+ATB AUTO**: train on the concatenation of Religious and ATB, test on the automatic segmentation.
7. **ATBWW:** train on the ATB whole words
8. **RELWW**: train on Religious whole words
9. **RELWW+ATBWW**: the concatenation of the training sets in 7 and 8.

3.2.1. POS Results and Discussion

Table 5(A) shows the results of the POS tagging experiments when tagging on segments, while Table 5(B) shows the results on whole words.

The first thing to notice in the results above is that the ATB-trained tagger performs poorly on religious Arabic. The difference in genre and the high ratio of out of vocabulary words are mainly to blame. While OOV words constitute 44% of the test set when training on the ATB, they are only

23% when training on the religious training set in spite of the fact that the ATB training set is 22 times as big (499884 versus 23001 words).

Experiment	Segment Accuracy	Word Accuracy
ATB GOLD	92.48%	82.82
ATB AUTO		78.62
REL GOLD	95.77%	90.55%
REL AUTO		87.33
REL(*10)+ATB GOLD	96.23	91.26
REL(*10)+ATB AUTO		88.22

Table 5(A): Segment-based POS results

There is also a considerable difference between tagging based on gold standard segmentation and that based on automatic segmentation. This hold true for all experiments, with a difference of 4.2% in the ATB experiment (82.82 vs. 78.62), 3.2% in the REL experiment (90.55 vs. 87.33), and 3% in the REL+ATB experiment (91.26 vs. 88.22). This shows that with more religious data available, the difference could shrink even more.

While segment-based tagging is prone to errors due to the problems resulting from segmentation, another approach is to use whole words with complex tags as units for tagging.

Experiment	Result
ATBWW	78.44%
RELWW	85.90
ATBWW+RELWW	86.96
ATBWW+RELWW*10 (rel train repeated 10 times)	87.24

Table 5(B): Whole word POS results

Results of whole word tagging show more or less the same patterns. The religious-trained tagger outperforms the ATB-trained tagger by 7.5%. The best results are obtained by the concatenation of the religious and ATB training data, repeating the earlier 10 times. This setting achieves an 8.8% absolute improvement over the ATB-trained tagger. This is only about 1% worse than the best-scoring automatic segment-based experiment, and we expect that with more data, the whole word approach would work better than with performing segmentation.

Whole word tagging results are impressive given that the ATB training set has 991 unique tags and the Religious training set has 569. The number of whole word tags in the test set is 324

3.2.2. POS Error Analysis

Due to the many experiments included, it may not be feasible to report on every error in every experiment. We will limit our error analysis to two experiments: ATB GOLD and REL GOLD. We will assume that in the two AUTO experiments, the extra errors are a result of erroneous segmentation.

Table 6 reports on the accuracies of the most common 20 tags in the test set. The top 20 tags count for 90% of all tags with NOUN ranking # 1 at 21.152%, the definite determiner DET # 2 at 11.3%, CONJ # 3 at 9.8%, prepositions PREP # 4 at 9.26 and PUNC # 5 at 8.24%. The worst scoring tags in the ATB experiment are ADJ, PV, NOUN_PROP, REL_PRON and NOUN, while the worst scoring ones in the REL experiment are ADJ, PV, NOUN_PROP, IV, and POSS_PRON_3MS.

Table 7 shows the confusion matrix between the three common low-scoring tags.

Tag	ATB Accuracy	REL Accuracy
NOUN	83.91%	92.08%
DET	99.81%	100%
CONJ	91.40%	100%
PREP	99.50%	98.50%
PUNC	93%	100%
NSUFF_FEM_SG	97.40%	99.35%
PV	71.77%	79.58%
IV	94.18%	88.38%
IV3MS	93%	99.61%
ADJ	66.94%	71.43%
SUB_CONJ	95.19%	99.03%
NEG_PART	100%	100%
PRON_3MS	98.63%	96.58%

POSS_PRON_3MS	94.12%	91.60%
NOUN_PROP	75.12%	82.05%
NUM	91.07%	96.43%
CASE_INDEF_ACC	97.24%	98.17%
REL_PRON	82.02%	94.38%] \
PRON_3FS	98.36%	98.36%
NSUFF_FEM_PL	100%	100%

Table 6: Frequent tag accuracies

Tag	ATB Confusions	REL Confusions
ADJ	NOUN 21.63% NOUN_PROP 11.48%	NOUN 25.31 NUM 1.63 NOUN_PROP 0.41
PV	NOUN 18% NOUN_PROP 7.8% PREP 1.2%	NOUN 12.91 IV 2.7 ADJ 1.8
NOUN_PROP	NOUN 19.66% ADJ 4.27%	NOUN 14.52% IV 0.85%

Table 7: Most common POS confusions

4. Related Work

To our knowledge, there exists no work that handles the morphological segmentation and part of speech tagging of religious Arabic, but some works are related which focused mostly on the Quran. Alhadj (2009) built a POS tagger for traditional Arabic with the ultimate aim of using the tagger for building a Quranic linguistic database. He trained his tagger on "Albayan-wa-tabyin" , a book by Al-Jahiz. However, the book is a literary one focusing on rhetoric, and the POS tagset used was very limited (13 tags). There is no clear evaluation of Quranic Arabic in the paper.

Another effort, also targeting the Quran, is that of the Quranic Arabic Corpus (corpus.quran.com) (Dukes and Buckwalter: 2010). The QAC is a comprehensive database of the Quran including morphological analysis, part of speech tagging, and dependency parsing. The Quranic Arabic Corpus differs from the work in this paper in that it is limited to the Quran, while we try to leverage a corpus and tools for many varieties of religious Arabic as attested by the selection of the three books in our tiny corpus. The POS tagset we use is generally more detailed than the one used in the QAC since we also segment and tag inflectional affixes, although their treatment of particles seems to be more appropriate, and we will try to include it in our future work.

Arabic POS tagging has long been an important topic in Arabic NLP in general, and several approaches exist. Habash and Rambow (2005) perform full morphological analysis that produces segmentation and POS tags as by-products. Mohamed and Kuebler (2010a, 2010b) and Kuebler and Mohamed (2011) treat segmentation as per letter classification task and perform POS tagging at the segment level where inflectional as well as syntactically functional tags are segmented and tagged. Diab et al (2007) and Diab (2009) use a pieplined approach in which they first perform tokenization then POS tagging using support vector machines without the use of a morphological analyzer. Kulick (2010) avoids the pieplined approach by performing simultaneous tokenization and POS tagging with a small tag set and reports promising results.

3 Conclusion

We have presented a small corpus of religious Arabic, and the results of word segmentation and POS tagging. We have compared the results obtained by training a segmenter and a POS tagger, and shown that even though the religious corpus is tiny, it produces better results than the ATB-trained segmenter and tagger. It is worth noting that even if we obtain a much larger newswire corpus for training, the results may not be better. We have checked the coverage in a 148,363,649 word portion of the Arabic Gigaword corpus (Graff et al, 2006) , and found that the OOV rate is 22.82% at the word type level and 9.32% at the token level.

Religious Arabic thus requires its own Treebank. We will work on adding more data to the current "tiny" selection making sure to cover the various aspects of religious Arabic as well as add more layers of annotation.

References

Mona Diab. 2009. Second generation tools (AMIRA 2.0): Fast and robust tokenization, pos tagging, and base phrase chunking. In Proceedings of 2nd International Conference on Arabic Language Resources and Tools (MEDAR), Cairo, Egypt, April.

Mona Diab, Kadri Hacioglu, and Daniel Jurafsky. 2007. Automatic processing of Modern Standard Arabic text. In Abdelhadi Soudi, Antal van den Bosch, and Gunter Neumann, editors, Arabic Computational Morphology, pages 159–179. Springer.

Kais Dukes and Timothy Buckwalter. 2010. A Dependency Treebank of the Quran using Traditional Arabic Grammar. In Proceedings of the 7th International Conference on Informatics and Systems (INFOS). Cairo, Egypt.

David Graff, Ke Chen, Junbo Kong, and Kazuaki Maeda (2006). The Arabic Gigaword Corpus. Second Edition. LDC Catalog No. LDC2006T02

Yahya O. Mohamed Elhadj. 2009. Statistical Part-of-Speech Tagger for Traditional Arabic Texts. Journal of Computer Science 5 (11): 794-800, 2009 . Science Publications.

Nizar Habash, Owen Rambow. 2005. Arabic tokenization, part-of-speech tagging and morphological disambiguation in one fell swoop. Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics. pp. 573-680.

Sandra Kuebler and Emad Mohamed. 2011. Part of speech tagging for Arabic. Natural Language Engineering.

Seth Kulick. 2010. Simultaneous tokenization and part-of-speech tagging for Arabic without a morphological analyzer. Proceeding of ACLShort '10 Proceedings of the ACL 2010 Conference Short Papers. Pages 342-347.

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

Emad Mohamed and Sandra Kübler. 2010b. Is Arabic part of speech tagging feasible without word segmentation? Proceedings of HLT-NAACL 2010, Los Angeles, CA.