Improving Arabic Tokenization and POS Tagging Using Morphological Analyzer

Michael N. Nawar

Department of Computer Engineering, Cairo University,
Cairo, Egypt
michael.nawar@eng.cu.edu.eg

Abstract. In this paper a new technique of tokenization and part-of-speech (POS) tagging for Arabic text is presented. The introduced technique uses the Arabic morphological analyzer to extract new features that will improve the stemming and the POS tagging. Applying standard evaluation metrics, the proposed tokenizer achieves an $F_{(\beta=1)}$ score of 99.99, and the POS tagger achieves an accuracy of 98.05%.

1 Introduction

Most of Natural language processing (NLP) systems such as information retrieval, text to speech, automatic translation and other use a part-of speech tagger for preprocessing. Supervised methods for part-of-speech (POS) tagging are expensive and time consuming as they depend on manually annotated data. However these methods achieve high results in NLP fields compared to unsupervised methods. Many of the Arabic words are ambiguous in their nature as tag of word can map to a noun, verb or adjective. It is believed that using a statistical approach which makes use of the morphological feature of the Arabic word would result in accurate, efficient and robust tagger that can be used in practical systems. Since both parsing and tagging Arabic words requires a stemming phase, a high accuracy in stemming phase implies a less accumulated error in further phases.

The basic idea of the proposed method is to recognize Arabic tokens and tagging them statistically using the Conditional Random Field learning approach by constructing a relevant model and feeding this model with some extra features extracted from the morphological analysis of each Arabic word. This concept is applied in the tokenization, normalization and POS tagging phase.

2 Arabic NLP and Data

There are three main categories of Arabic language; classical the language of Quran, modern standard (MSA) which is a simplified form of classical that is extracted from news and written documents, and dialectical Arabic which differs from one country to another. One variation of it is the colloquial language which is the daily used language by Egyptians.

A.E. Hassanien et al. (Eds.): AMLTA 2014, CCIS 488, pp. 46–53, 2014.
© Springer International Publishing Switzerland 2014

In general Arabic has a very rich morphological language where each word can include number, gender, aspect, case, mood, voice, mood, person, and state. The Arabic basic word form can be attached to a set of clitics representing object pronouns, possessive pronouns, particles and single letter conjunctions. Obviously the previous features of Arabic word increase its ambiguity. Generally Arabic stems can be attached three types of clitics orderd in their closeness to the stem according to the following formula:

$$\{[proclitic1]\{[proclitic2]\{Stem[Affix][Enclitic]\}\}\}$$

Where proclitic1 is the highest level clitics that represent conjunctions and is attached at the beginning such as the conjunction [(و, w, and), (ف, f, then)]. Proclitic2 represent particles [(ب, b, with/in), (ل, l, to/for); (ك, k, as/such)]. Enclitics represent pronominal clitics and are attached to the stem directly or to the affix such as pronoun [(ه, h , his), (هم, hm , their/them)].

The following is an example of the different morphological segments in the word that has the stem (قدر, qdr ,power), the proclitic conjunction (و, w, and) , the proclitic particle (ب, b ,with/in) , the affix (ات, At ,for plural) ,and the cliticized pronoun (ه, h , his). The set of proclitics considered in this work are the particles prepositions b, l, k, meaning by/with, to, as respectively, the conjunctions w, f, meaning and, then respectively. Arabic words may have a conjunction and a preposition and a determiner cliticizing to the beginning of a word. The set of possible enclitics comprises the pronouns and (possessive pronouns) y, nA, k, kmA, km, knA, kn, h, hA,hmA, hnA, hm, hn, respectively, my (mine), our (ours), your (yours), your (yours) [masc. dual], your (yours) [masc. pl.], your (yours) [fem. dual], your (yours) [fem.pl.], him (his), her (hers), their (theirs) [masc. dual], their (theirs) [fem. dual], their (theirs) [masc. pl], their (theirs) [fem. pl.]. An Arabic word may only have a single enclitic at the end. A token is defined as a (stem + affixes), proclitics, enclitics, or punctuation.

The data used for training and testing the stemmer and the POS tagger is the Arabic Treebank part 1 [1] which consists of 734 news articles (140kwords corresponding to 168k tokens after semi-automatic segmentation) covering various topics such as sports, politics, news, etc.

3 Related Work

A lot of the existing systems tend to target a specific application or a POS tag set that is not general enough for different applications. For example Shereen Khoja in (2001) [10] reports preliminary results on a hybrid, statistical and rule based, POS tagger, APT. APT yields 90% accuracy on a tag set of 131 tags including both POS and inflection morphology information. Diab et al. (2007) [1] perform a large-scale corpus-based evaluation of their approach. They use Yamcha SVM classifier based learner for three different tagging tasks: word tokenization, POS tagging and base phrase chunking with a collapsed tag set achieving a $F_{(\beta=1)}$ score of 99.12 on word tokenization and an accuracy of 96.6%

on POS tagging respectively. Diab (2009) [7] extended the work on Diab et al. (2007) to multiple tag set, instead of the PATB (Penn. Arabic Treebank) reduced tag set. Habash and Rambow (2005) [2] use SVM classier for individual morphological features and an ad-hoc combining scheme for choosing among competing analysis achieving an accuracy of 97.5%. Mansour (2007) [6] port an HMM Hebrew tagger to Arabic yielding to an accuracy of 96.1% for POS tagging. AlGahtani et al. (2009) [4] use transition based learning for the task of POS tagging, achieving an accuracy of 96.9%. Kulick, S. (2010) [5] performs simultaneous tokenization and POS tagging without a morphological analyzer, achieving an accuracy of 95.1% for POS tagging.

It is not a simple matter to compare results with previous work, due to differing evaluation techniques, data sets, and POS tag sets. In this paper, the results are compared with Diab et al. (2007) (SVM system) and Habash and Rambow (2005) (Majority system); because both of those papers and both of them work on the same range of data PATB (Penn. Arabic Treebank) part1, they report the results based on the PATB reduced tag set, they assume gold tokenization for evaluation of POS results, and the main concern is to report the highest accuracy unlike AlGahtani et al. (2009) and Kulick, S. (2010) where their main concern is the speedup.

4 Tokenization Phase

In this phase, the classifier takes an input of raw text, without any processing, and assigns each character the appropriate tag from the following tag set B-PRE1, B-PRE2, B-WRD, I-WRD, B-SUFF, I-SUFF. Where I denotes inside a segment, B denotes beginning of a segment, PRE1 and PRE2 are proclitic tags, SUFF is an enclitic, and WRD is the stem plus any affixes and/or the determiner Al. Two experiments have been conducted to achieve the final tokenizer: base line and binary feature experiments. The base line experiment is used to check the effect of using a CRF classifier instead of a SVM classifier in the task of tokenization. In the binary feature experiment a new feature has been proposed in addition to the features used in the base line experiment, and the effect of the binary feature in the task of tokenization is checked.

4.1 Baseline Experiment (CRF-TOK)

This experiment is based on the experiment of (Diab et al., 2007) but instead of using SVM classifier the CRF suite classifier is used. The classifier training and testing data is characterized as follows:

- Input: A sequence of transliterated Arabic characters processed from left-to-right with break markers for word boundaries.
- Context: A fixed-size window of -5/+5 characters centered at the character in focus.
- Features: All characters and previous tag decisions within the context.

4.2 Binary Feature Experiment (BF-TOK)

A new feature is proposed in this experiment and this feature is added to the feature set in the baseline experiment. BAMA-v2.0 (Buckwalter Arabic morphological analyzer version 2.0) is used to define a binary feature of length 6 where each bit in the feature is mapped to one of the 6 tags in the tokenization tag set. A bit is set if at least one analysis in the morphological analyses of the word, the character is assigned the tag corresponding to the bit.

For example the word (وحيد, wHyd) has two possible tokenization schemes: (و+حيد, w+Hyd) or (وحيد, wHyd); then (و, w) could be (B-PRE1 or B-WRD) then in the binary feature of the character there will be 2 bits set which map to B-PRE1 and B-WRD, (ح, H) could be (B-WRD or I-WRD) then in the binary feature of the character there will be 2 bits set which map to B-WRD and I-WRD, (ي, y) and (د, d) could be only (I-WRD) then in the binary feature of the characters there will be only one bit set which map to I-WRD. Table (1) shows the binary feature of each character of the word (وحيد, wHyd).

Table 1. Tokenization Binary Feature

Arabic Letter	Transliterated Letter	Binary Feature					
		B-PRE1	B-PRE2	B-WRD	I-WRD	B-SUFF	I-SUFF
و	w	1	0	1	0	0	0
ح	H	0	0	1	1	0	0
ي	y	0	0	0	1	0	0
د	d	0	0	0	1	0	0

If the word is not analyzed by the morphological analyzer (out of vocabulary); then all 7 bits of the binary feature will be set.

5 POS Tagging Phase

In this phase, the classifier takes an input of tokenized text, and it assigns each token an appropriate POS tag from the Arabic Treebank collapsed POS tags, which comprises 24 tags as follows: ABBREV, CC, CD, CONJ+NEG PART, DT, FW, IN, JJ, NN, NNP, NNPS, NNS, NO FUNC, NUMERIC_COMMA, PRP, PRP$, PUNC, RB, UH, VBD, VBN, VBP, WP, WRB}. Two experiments have been conducted to achieve the final POS tagger. The first experiment is used to check the effect of using a CRF classifier instead of a SVM classifier in the task of tokenization. In the second, the binary feature experiment a new feature has been proposed in addition to the features used in the base line experiment, and the effect of the binary feature in the task of POS tagging is checked.

5.1 Base Line Experiment (CRF-POS)

This experiment is based on the experiment of (Diab et al., 2007) but instead of using SVM classifier a CRF classifier is used. The classifier training and testing data is characterized as follows:

- Input: A sequence of transliterated Arabic tokens processed from left-to-right with break markers for word boundaries.
- Context: A window of -2/+2 tokens centered at the focus token.
- Features: Every character N-gram, N¡=4 that occurs in the focus token, the 5 tokens themselves, POS tag decisions for previous tokens within context.

5.2 Binary Feature Experiment (BF-POS)

A new feature is proposed in this experiment and this feature is added to the feature set in the baseline experiment. BAMA-v2.0 (Buckwalter Arabic morphological analyzer version 2.0) is used to define a binary feature of length 24 where each bit in the feature is mapped to one of the 24 tags in the collapsed POS tag set. A bit is set when its corresponding tag exists in the morphological analysis of a token.

For example the word (كتب, ktb) has 3 different reduced POS tags: VBD then it will mean (write), VBN then it will mean (be written), and NN then it will men (book); so there will be 3 bits set to one in the binary feature of the (كتب, ktb) word corresponding to VBD, VBN and NN. While you can find a word like (الولد, Alwld) has only one reduce POS tag which is NN and it have only one meaning the boy. In table (2), you can find the binary feature for the words of the sentence (كتب الولد الدرس, ktb Alwld Aldrs, The boy wrote the lesson).

Table 2. POS Tagging Binary Feature

Arabic Word	Transliterated Word	Binary Feature					
		VBD	VBN	NN	JJ	NNS	...
كتب	ktb	1	1	1	0	0	0
الولد	Alwld	0	0	1	0	0	0
الدرس	Aldrs	0	0	1	0	0	0

But for the word (يكتب, yktb) it has only one reduced POS tag: VBP which means (write); so there will be only one bit set in the binary feature which map to VBP. If the word is not analyzed by the morphological analyzer (out of vocabulary) like the word (الفالوجة, AlfAlwjp) which is a village in Palestine, then there will be 5 bits set in the binary feature which map to JJ, NN, NNS, NNP, and NNPS.

6 Empirical Results

For the evaluation of these experiments, k-fold algorithm was used by setting the parameter k to five so the Penn Arabic tree bank part1 is randomly partitioned into five portions of equal size. In each iteration of the k- fold algorithm four portions were used for training the model and one portion was used for testing the model. The cross-validation process is then repeated five times (the folds), with each of the k subsamples used exactly once as the testing data. The five results from the folds were averaged to produce the model evaluation. This evaluation scheme was applied for both the tokenization and POS tagging. Then the following performance measures are calculated for each experiment

$$macro\ average\ precision = \frac{1}{n}\sum_{i=1}^{n} precision(tag(i))$$

$$macro\ average\ recall = \frac{1}{n}\sum_{i=1}^{n} recall(tag(i))$$

$$macro\ average\ F_{(\beta=1)} = \frac{1}{n}\sum_{i=1}^{n} F_{(\beta=1)}(tag(i))$$

$$Accuracy = \frac{number\ of\ true\ results}{number\ of\ true\ and\ false\ results}$$

Then the proposed method is compared with the SVM based approach [1] and the Majority system [2]. The comparison between the proposed method and the SVM approach and the majority system will be in the accuracy and the $F_(= 1)$ of the tokenizer and in the accuracy of the POS tagger, because these are the only performance measures they have reported. The tool used for evaluation is the evaluation tool in the CRF- Suite software package.

6.1 Tokenization Phase Evaluation

Table (3) compares the different experiments applied to the Tokenization task where the row represents the experiment and the column represents the macro average performance measure.

Table 3. Tokenization Phase Evaluation Results

	Precision	Recall	$F_{\beta=1}$	Accuracy	Error
CRF-TOK	0.99835	0.99926	0.99880	99.98%	0.02%
BF-TOK	0.99998	0.99908	0.99952	99.99%	0.01%

The performance of BF-TOK is almost perfect. Comparing BF-TOK to other Arabic tokenizers like: SVM-TOK which has an accuracy of 99.77% and an F score of 99.12; and with the Majority-TOK which has an accuracy of 99.3% and an $F_{\beta=1}$ of 99.1; the improved stemmer reduces the error by about 95.65% compared to the SVM-TOK, and by 98.57% from the Majority system tokenizer.

6.2 POS Tagging Phase Evaluation

Table (4) compares the different experiments applied to the POS tagging task where the row represents the experiment and the column represents the macro average performance measures.

Table 4. POS Tagging Phase Evaluation Results

	Precision	Recall	$F_{\beta=1}$	Accuracy	Error
CRF-POS	0.83279	0.77210	0.79130	96.10%	3.9%
BF-POS	0.84872	0.81236	0.82695	98.05%	1.95%

The BF-POS is compared with other Arabic POS taggers like: SVM-POS which has an accuracy of 96.6%, and the Majority-POS which has an accuracy of 97.6%. The result was that the proposed POS tagger reduces the error by 42.65% compared to the SVM-POS tagger and by 18.75% compared to the Majority POS tagger.

7 Conclusion and Future Work

In this research, the morphological analyzer is introduced to improve stemmer and POS tagger. Using the benchmark data set improvements in both tokenization and POS stages have been reached. First the CRF classifier is used instead of SVM. This resulted in an error reduction by 91.30% in the tokenization stage. Then the new binary feature (BF) extracted from the morphological analyses of the word is added to the feature set. This binary feature is language independent and highly accurate. It resulted in an error reduction by 95.65% and 18.75% in the tokenization and POS stage, respectively.

To achieve the targeted improvement the proposed system needs extra processing for the extraction of the binary feature. This extra processing could be minimized by using caching techniques in the implementation of the task of binary feature (BF) extraction.

There are numerous ways to extend this research work. The proposed binary feature BF will be tested on other languages like English. In addition, the performance of the Arabic POS tagging system additional features will be developed to further improve the performance. Last but not least, a wider context and more data will be used for testing.

References

1. Diab, M., Hacioglu, K., Jurafsky, D.: Automated methods for processing Arabic text: From tokenization to base phrase chunking. In: van den Bosch, A., Soudi, A. (eds.) Arabic Computational Morphology: Knowledge-based and Empirical Methods. Kluwer/Springer (2007)

2. Habash, N., Rambow, O.: Arabic Tokenization, Part-of-Speech Tagging and Morphological Disambiguation in One Fell Swoop. In: Proc. of the American Association of Computational Linguistic Conference (ACL) Short Papers, Michigan, USA (2005)
3. Habash, N., Rambow, O.: Morphological analysis and generation for Arabic dialects. In: Proc. of the Workshop on Computational Approaches to Semitic Languages in the American Association of Computational Linguistic Conference (ACL), Michigan, USA (2005)
4. AlGahtani, S., Black, W., McNaught, J.: Arabic Part-of-Speech Tagging Using Transformation-Based Learning. In: Proc. of the Second International Conference on Arabic Language Resources and Tools, Cairo, Egypt (April 2009)
5. Kulick, S.: Simultaneous Tokenization and Part-of-Speech Tagging for Arabic without a Morphological Analyzer. In: Proc. of the American Association of Computational Linguistic (ACL) Conference Short Papers, Uppsala, Sweden (July 2010)
6. Mansour, S., Sima'an, K., Winter, Y.: Smoothing a Lexicon-based POS tagger for Arabic and Hebrew. In: Proc. of the American Association of Computational Linguistic Workshop on Computational Approaches to Semitic Languages: Common Issues and Resources, Prague, Czech Republic (2007)
7. Diab, M.: Second generation tools (AMIRA 2.0): Fast and robust tokenization, pos tagging, and base phrase chunking. In: Proc. of 2nd International Conference on Arabic Language Resources and Tools (MEDAR), Cairo, Egypt (April 2009)
8. Maamouri, M., Bies, A., Buckwalter, T.: The penn arabic treebank: Building a largescale annotated arabic corpus. In: Proc. of NEMLAR Conference on Arabic Language Resources and Tools, Cairo, Egypt (2004)
9. Tamah, E., Al-Shammari, J.L.: Towards an Error-Free Arabic Stemming. In: Proc. of the American Association of Computational Linguistic (ACL) Conference on Information and Knowledge Management, New York, NY, USA (2008)
10. Khoja, S., Garside, P., Knowles, G.: A tagset for the morphosynactic tagging of Arabic. In: Proc. of Corpus Linguistics. Lancaster University, Lancaster (2001)