Arabic Tokenization, Part-of-Speech Tagging and Morphological Disambiguation in One Fell Swoop

Nizar Habash and Owen Rambow
Center for Computational Learning Systems
Columbia University
New York, NY 10115, USA

By

Hussain AL-Ibrahem

214131

Outline

- Introduction
- General approach
- Preparing The Data.
- Classifiers for linguistic Features
- Choosing an Analysis
- Evaluating Tokenization
- Conclusion

Introduction

- The morphological analysis of a word consists of determining the values of a large number of (orthogonal) features, such as basic part-of-speech (i.e.,noun, verb, and so on), voice, gender, number, information about the clitics.
- Arabic gives around 333000 possible completely specified morphological analysis.
- In first 280000 words in ATB 2200 morphological tags used.
- English have about 50 tags cover all.

Introduction

- **morphological disambiguation** of a word in context, cannot be done using method for English because of data sparseness.
- Hajic (2000) show that morphological disambiguation can be aided by morphological analyzer (given a word without syntax give all possible tags).

General approach

- Arabic words are often ambiguous in their morphological analysis. This is due to Arabic's rich system of affixation and clitics.
- On average, a word form in the ATB has about 2 morphological analyses.

Example

#	lexeme	gloss	POS	Conj	Part	Pron	Det	Gen	Num	Per	Voice	Asp
2	<ilaY	and to me	P	YES	NO	YES	NA	NA	NA	NA	NA	NA
3	waliy	and I follow	V	YES	NO	NO	NA	neut	sg	1	act	imp
5	|liy~	and automatic	AJ	YES	NO	NO	NO	masc	sg	3	NA	NA

Figure 1: Possible analyses for the word والي *wAly*

General Approach

- In this approach tokenizing and morphological tagging are the same operation which consist of 3 phases:
  1- Preparing The Data.
  2- Classifiers for linguistic Features.
  3- Choosing an Analysis.

Preparing The Data

- The data used came from Penn Arabic Tree bank and the corpus is collected from news text.
- The ATB is an ongoing effort which is being released incrementally.
- The first two releases of the ATB has been used ATB1 and ATB2 which are drawn from different news sources.

Preparing The Data

- ATB1 and ATB2 divided into development, training and test corpora with 12000 word token in development and test one and 120000 word in training corpora.
- ALMORGEANA morphological analyzer used the database from Buckwalter Arabic morphological analyzer BUT in analysis mode produce an output in the lexeme and feature format rather than steam-and-affix format.

Preparing The Data

- Training data consist of a set of all possible morphological analysis for each word with unique correct analysis marked and it is on ALMORGEAN output format.
- To obtain the data we have to match data in ATB to the lexeme-and-feature represnted by almorgean and this matching need some heuristic since representations are not in ATB.

Example & Notes

- Word(نحو) will be tagged as AV,N or V.
- Show that from 400 words chosen randomly from TR1 & TR2, 8 cases POS tagging differ than ATB file.
- One case of 8 was plausible among N , Adj , Adv and PN resulting of missing entries in Buckwalter lexicon.
- The other 7 filed because of handling of broken plural at lexeme level.

Preparing The Data

- From the before numbers the data representation provide an adequate basis for performing machine learning experiments.

Unanalyzed words

- Words that receive no analysis from the morphological analyzer.
- Usually proper nouns.
- (برلوسكوني) which does not exist in Buckwalter lexicon BUT ALMORGAN give 41 possible analyses including a single masculine PN.
- In TR1 22 words are not analyzed because Buckwalter lexicon develop in it.
- In TR2 737 (0.61%) words without analysis.

Preparing The Data

- In TR1 (138,756 words) ➔ 3,088 NO_FUNC POS labels (2.2%).
- In TR2 (168296 words) ➔ 853 NO_FUNC (0.5%).
- NO_FUNC like any POS tag but it is unclear in the meaning.

Classifiers for linguistic Features

Morphological Feature

Feature Name	Description	Possible Values	POS that Carry Feature	Default
POS	Basic part-of-speech	See Footnote 9	all	X
Conj	Is there a cliticized conjunction?	YES, NO	all	NO
Part	Is there a cliticized particle?	YES, NO	all	NO
Pron	Is there a pronominal clitic?	YES, NO	V, N, PN, AJ, P, Q	NO
Det	Is there a cliticized definite determiner +ال *Al*+?	YES, NO	N, PN, AJ	NO
Gen	Gender (intrinsic or by agreement)	masc(uline), fem(inine), neut(er)	V, N, PN, AJ, PRO, REL, D	masc
Num	Number	sg (singular), du(al), pl(ural)	V, N, PN, AJ, PRO, REL, D	sg
Per	Person	1, 2, 3	V, N, PN, PRO	3
Voice	Voice	act(ive), pass(ive)	V	act
Asp	Aspect	imp(erfective), perf(ective), imperative	V	perf

Classifiers for linguistic Features

- As training features tow sets used. These sets are based on the Morphological Feature and four hidden for which do not train classifiers.
- Because they are returned by the morphological analyzer when marked overtly in orthography but not disambiguates.
- These features are indefiniteness, idafa (possessed), case and mood.

Classifiers for linguistic Features

- For each 14 morphological features and possible value a binary machine defined which give us 58 machine per words
- Define Second set of features which are abstract over the first set state whether any morphological analysis for that word has a value other than ‘NA’. This yields a further 11 machine learning
- 3 morphological features never have the value ‘NA.
- two dynamic features are used, namely the classification made for the preceding two words.

Classifiers for linguistic Features

Method		Class	BL	Class
Test		TE1	TE2	TE2
POS		97.7	91.1	95.5
Conj		99.9	99.7	99.9
Part		99.9	99.5	99.7
Pron		99.6	98.8	99.0
Det		99.2	96.8	98.3
Gen		99.2	95.8	98.2
Num		99.4	96.8	98.8
Per		98.7	94.8	98.1
Voice		99.3	97.5	99.0
Asp		99.4	97.4	99.1

BL : baseline

Choosing an Analysis.

- Once we have the results from the classifiers for the ten morphological features, we combine them to choose an analysis from among those returned by the morphological analyzer .
- two numbers for each analysis. First, the agreement is the number of classifiers agreeing with the analysis. Second, the weighted agreement is the sum, over all classifiers of the classification confidence measure of that value that agrees with the analysis.

Choosing an Analysis.

We use Ripper (Rip) to determine whether an analysis from the morphological analyzer is a "good" or a "bad" analysis.

We use the following features for training: we state whether or not the value chosen by its classifier agrees with the analysis, and with what confidence level. In addition, we use the word form. (The reason we use Ripper here is because it allows us to learn lower bounds for the confidence score features, which are real-valued.) In training, only the correct analysis is good. If exactly one analysis is classified as good, we choose that, otherwise we use Maj to choose.

Corpus	TE1		TE2	
Method	All	Words	All	Words
BL	92.1	90.2	87.3	85.3
Maj	96.6	95.8	94.1	93.2
Con	89.9	87.6	88.9	87.2
Add	91.6	89.7	90.7	89.2
Mul	96.5	95.6	94.3	93.4
Rip	96.2	95.3	94.8	94.0

classifiers are trained on TR1; in addition, Rip is trained on the output of these classifiers on TR2.

Choosing an Analysis.

- The difference in performance between TE1 and TE2 shows the difference between the ATB1 and ATB2
- the results for Rip show that retraining the Rip classifier on a new corpus can improve the results, without the need for retraining all ten classifiers

Corpus	TE1		TE2	
Method	All	Words	All	Words
BL	92.1	90.2	87.3	85.3
Maj	96.6	95.8	94.1	93.2
Con	89.9	87.6	88.9	87.2
Add	91.6	89.7	90.7	89.2
Mul	96.5	95.6	94.3	93.4
Rip	96.2	95.3	94.8	94.0

Evaluating Tokenization

- The ATB starts with a simple tokenization, and then splits the word into four fields: conjunctions; particles; the word stem; and pronouns. The ATB does not tokenize the definite article +Al+.
- For evaluation, we only choose the Maj chooser, as it performed best on TE1.
- First evaluation, we determine for each simple input word whether the tokenization is correct and report the percentage of words which are correctly tokenized

Evaluating Tokenization

- In the second evaluation, we report on the number of output tokens. Each word is divided into exactly four token fields, which can be either filled or empty or correct or incorrect.
- report accuracy over all token fields for all words in the test corpus, as well as recall, precision, and f-measure for the non-null token fields
- The baseline BL is the tokenization associated with the morphological analysis most frequently chosen for the input word in training.

Conclusion

- Preparing The Data.
- Classifiers for linguistic Features
- Choosing an Analysis
- Evaluating Tokenization

Q&A