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

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.

General approach:

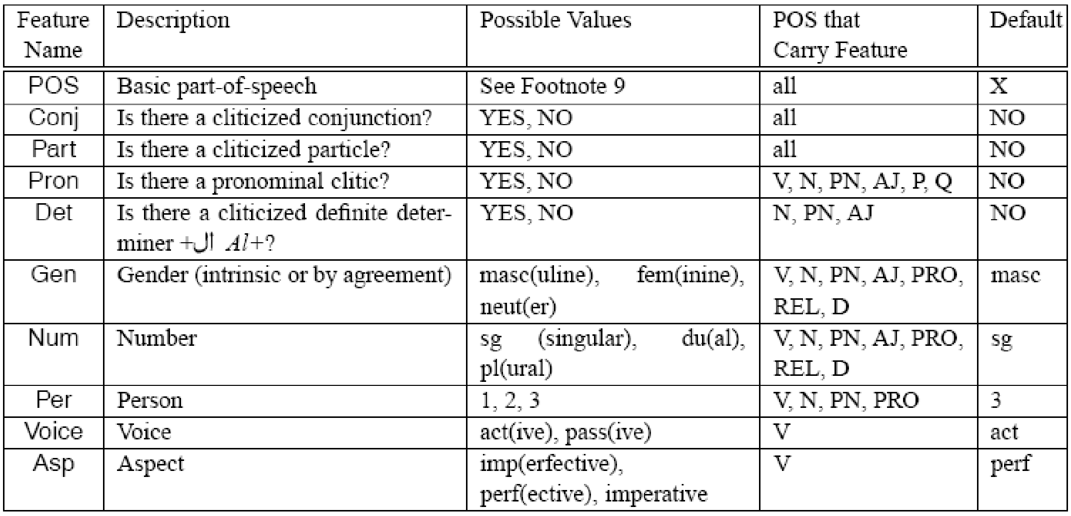
Arabic words are often ambiguous in their morphological analysis. This is due to Arabic's rich system of affixation and clitics. On average, a word form in the ATB has about 2 morphological analyses.

- **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 first two releases of the ATB has been used ATB1 and ATB2 which are drawn from different news sources.Unanalyzed words Words that receive no analysis from the morphological analyzer.

Classifiers for linguistic Features



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

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.

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+. First evaluation, we determine for each simple input word whether the tokenization is correct and report the percentage of words which are correctly tokenized 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.

Obstacles:

- **Language used in paper is high level language.**
- **No explanation of calculation.**
- **Some terms are not defined in paper.**

Skills and benefits:

- **Present a topic in front of some people.**
- **Prepare presentation.**
- **how to use a morphological analyzer for tokenization, part-of-speech tagging, and morphological disambiguation in Arabic.**
- **that the use of a morphological analyzer is beneficial in POS tagging.**

True False Questions:

1. **Unanalyzed words Words that receive no analysis from the morphological analyzer. T**
2. **the weighted agreement is the sum, over all classifiers of the classification confidence measure of that value that agrees with the analysis. T**
3. **In Evaluating Tokenization phase in second evaluation Each word is divided into exactly 2 token fields. F**