

Automatic Part-of-Speech Tagging

of Arabic Text

العَنْوُكَةُ الْآلِيَّةُ لِتَصْرُوحِ اللُّغَةِ الْعَرَبِيَّةِ



UNIVERSITY OF LEEDS

School of Computing,
FACULTY OF ENGINEERING

Majdi Sawalha

Supervisor: Dr. Eric Atwell

INTRODUCTION

Part of Speech Tagging: is the process of assigning a grammatical part of speech tag to each word based on its context.

A **tag:** is a code which represents some features or set of features and is attached to the word in a text. Tags may carry single or complex information.

The **tagset** may represent morphological, syntactic and/or phrase structure information.

Tagging Applications

- A good tagger can serve as a preprocessor where more abstract levels of analyses benefit from reliable low-level information.
- Large tagged text corpora are used as data for linguistic studies.
- Information technology applications; e.g. text indexing and retrieval can benefit from Part-of-Speech information.
- Speech processing.

RESEARCH QUESTIONS

How to widen the scope of Arabic Part-of-Speech tagging, to develop a system which can process Arabic text in wide range of formats, domains, and genres of both vowelized and non-vowelized text?



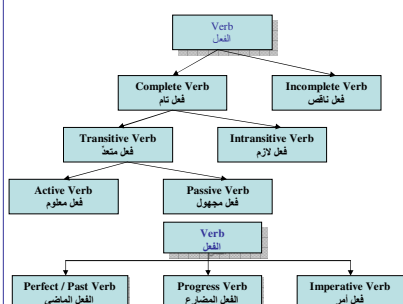
Can richer lexical resources derived from dictionaries and grammar text books improve the coverage of morphological analysis for wider range of Arabic text formats, domains and genres?

How do we evaluate existing Part-of-Speech taggers and new Part-of-Speech taggers on a wider range of text formats, domains, genres, and vowelized and non-vowelized text?

How do I make the best reuse of existing tagger components and methods?

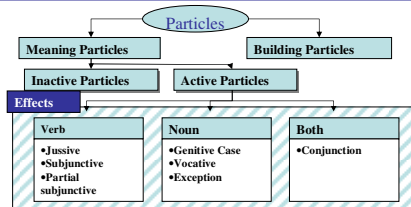
ARABIC WORD CLASSIFICATION

Arabic language linguists classify words in Arabic into 3 main categories. **Verbs, Nouns and Particles.**



Nouns

Arabic language linguists distinguish between 21 types of nouns including **Adjectives, Adverbs, Personal Pronouns, ... etc.**



EVALUATING ARABIC TAGSETS

- Every researcher has developed a tagset.
- A comparison of different tagsets will show
 - The **number** of tags used,
 - The **purpose** of using the tagset.
 - The **source of information** when designing the tagset.
 - The **errors in classifying** tags into their categories.
- Combining tagsets into one **reliable and multi-level tagset** that varies from **minimal tagset** to **more detailed one**.

ARABIC LANGUAGE CHALLENGES

GOLD STANDARD FOR EVALUATION

1 Tokenization, Agglutination and Complex Morphology
Words can be prefixed or suffixed with clitics. Clitics can be concatenated one after the other. A single word can comprise up to four independent morphemes.

2 Vowels & Diacritic Marks: Arabic has 3 long vowels (Alif , waw , yaa) and 3 short vowels (Fatha , Damma , Kasra), Sukun and Shaddah

3 Grammatical Ambiguity : Lexical forms with vowel marks has grammatical ambiguity rate of 2.8 on average. This rate increases by the absence of vowels to reach 5.6 possible tags per lexical form.

• Different text domains, formats and genres of both vowelised and non-vowelised text.

- The **Qur'an**.
- **Newspaper** text.
- **Magazines**.
- **School books**.
- **Children's books**.
- **Blogs**

• Gold Standard will be checked by Arabic language scholars.

الم أحسب التمس أن يتركوا أن يقولوا آمنا وهم لا يفتنون ولقد فتنا الذين من قبلهم فليعلمن الله الذين صدقوا وليعلمن الكافرين أم حسب الذين يعلمون السيات أن يسبقونا ساء ما يحكمون من كان يرجو لقاء الله فإن أجل اللوات وهن عن العالمين والذين آمنوا و عملوا الصالحات لذكرون عنهم سياتهم ولنجزينهم أحسن الذي كانوا يعملون ووصينا الإنسان بوالديه حسنا وإن جاهدك بشرك به يا ليس لك به علم فلا تطعهما إلى مرجعكم فأنتم بما كنتم تعملون والذين آمنوا و عملوا الصالحات لندخلنهم في الصالحين

سنتي العولة وإلى وقت عند متيرة للأسئلة والأجوبة وفي هذا المقال ولقة تأمل عميقة في بعض هذه الأسئلة بدأت منذ فترة موجه جديدة من الكتابات تروج للعولة باعتبارها الشكل الجديد لحياة البشر في ظل القطب الأمريكي وهناك غط من هذه الكتابات يروج للمنطق الأمريكي متعدد الأعراف والثقافات بوصفه النمط الأمثل للحياة في القرية الكونية الجديدة التي قاربت وسائل الاتصالات والمواصلات ونظم المعلومات ووسائل الإعلام بين أجزائه المختلفة ويشتر أصحاب هذه النظرة بشر نوح جديد بشر كوزموبوليون

BROAD LEXICAL RESOURCE | EXISTING TAGGERS | HYBRID POS Tagger

- 15 Arabic language dictionaries are used
- The lexicon contains:
 - **Roots and single words.**
 - **Multi-word expressions.**
 - **Idioms.**
 - **Collocations** requiring special part of speech tags.
 - **Words with special part of speech tags.**
 - **Meanings.**



- **Evaluating existing Part-of-Speech tagger components.**
 - Gold Standard - Fair measurements
 - Multi-level tagset
- **Analyzing & re-implementing algorithms of Part-of-Speech taggers.**
 - Best tagger components need to be re-implemented, using Python.
 - Python will simplify the integration of the Part-of-Speech tagger into the NLTK

- Novel algorithm leading to **hybrid Part-of-Speech tagger** for Arabic text which combines **best components of existing taggers** with **novel resources and components.**
 - Integrating **best tagger components** together.
 - Integrating **Prior-knowledge lexical resource.**
 - Integrating **Morphological analyser.**
 - Using **unsupervised learning algorithms** to solve the problem of **unknown words.**