Part-of-Speech Tagging for Arabic language

Omar al-qaisi, Marwan al-ramahi, Motassem Naqawah

1-Introduction

Part-of-speech (POS) tagging is one of the most important addressed areas in the natural language processing (NLP). There are effective POS taggers for many languages. We tried to develop a POS tagger for the Arabic language, specifically for the modern standard Arabic (MSA), because it's the language used in the formal textbooks and news.

The objective of our solution is to firstly create a tokenizer that splits any file you choose into a list of words with removing any punctuations and numbers from the list. And secondly create a POS which takes the list of words from tokenizer and then tag each word with its appropriate POS(verb, noun, particle) based on a combination of rules. Finally we created a golden corpus from a sample of the actual corpus folder to test our algorithm and see how accurate and precise with its tagging.

2-Related work

We searched for a lot of resources on the internet to help us build our own POS, and we found a plenty of reports and research on this subject, that was useful for us we like to mention.

- ### Part-of-Speech Tagging for Arabic Gulf Dialect Using Bi-LSTM

People worked on it : Randah Alharbi, Walid Magdy , Kareem Darwish , Ahmed AbdelAli , Hamdy Mubarak.

Date: MAR 2018

They mentioned and explained how a one Arabic word that can have different pronunciations, meanings and different part-of-speech and how there are some words in Arabic complex segmented words like وسنجعلكم.

They present a more effective POS tagger for the Arabic Gulf dialect than currently available Arabic POS taggers. And they focused on the Dialect Arabic(DA) which is the language used in the social media and in the day to day basis between people.

Link: https://www.aclweb.org/anthology/L18-1620.pdf

- **A Review of Part of Speech Tagger for Arabic Language**

People worked on it : Salem Salameh

Date: June 2018

The aim of his paper is to review the implementation of Part of Speech (POS) Tagger for Arabic Language which will help in building accurate corpus for Arabic Language.

The POS tagger has been implemented using different methods for Arabic languages like Rule-Based Model, Statistical Model and Neural Network Tagger. In addition, there are another methods such hybrid system and decision trees Besides, other methods are used like transformation-based.

Link:
https://www.researchgate.net/publication/326464406_A_Review_of_Part_of_Speech_Tagger_for_Arabic_Language

3-Tagger approach

We used two main approaches to get the best accuracy we can achieve from tagging the text:

1-Rule Based approach: this was our main approach because its reliable and more efficient than the patterns approach.

2-pattern approach: we used it as a complimentary approach to the Rule Based, to reach higher accuracy, and because in the rule based its hard to cover all the rules in Arabic language.

4-Tag Set

Rule based:

- Conjunctions that have a noun after it:صباح ,حين , دون , ليت, و, لعل, كأن, إنَّ, تحت, وراء,حيث في, من ,عن,إلى , على , يا, كل, فوق,أمام ,ظهر,
- Conjunctions that have a verb after it: أن, لن , كي, حتى, لم, لما, لا, إن, ما, مهما, متى,أينما, حيثما
- kan nouns: كان, صار, أصبح, أضحى, ظل, أمسى, بات, ليس
- nouns that starts with(noun prefix) : ال , لل, فال, كال,بال
- verbs that starts with(verb prefix):سي, سن, ست, سأ

- verb patterns : استا*ت**, **ا**وع
- noun patterns: ***م*, *ا**م, ة*ا**, ة***م, ى***,ي**