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 tagger 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

- ## Arabic Word Segmentation

People worked on it : Yassine Benajiba, Imed Zitouni

Link: https://pdfs.semanticscholar.org/55b5/c93ad5035740790f0d7e364d2c065f04fb6e.pdf

Due to its Semitic origins the Arabic language uses 3 main things to form the words.

1-**Derivation:**is used to form a certain meaning using the root and a template. For instance the word مكتوب is derived from the root كتب using the template مفعول .

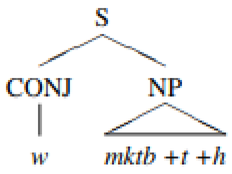
2-**Inflection**: By using inflection, one is able to obtain, for instance, the feminine form of this word, i.e. مكتوبة or its plural form, i.e. مكتوبون

3- **Enclitization**: consists of adding prefixes and suffixes to the words in order to obtain further meaning. For instance, if we wanted to express "and written", then we want to add the conjunction as a the prefix و to the word written which we have introduced previously and thus we would obtain ومكتوب . It is, however, important to mention that the real world data exhibit much more complicated cases using more than one prefix and having both suffixes added during inflection and enclitization.

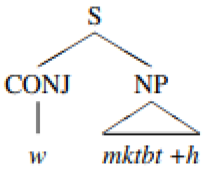
Arabic Segmentation Schemes

The most widely adopted segmentation schemes for natural language processing tasks are:

- **Morphological Segmentation :** aims at segmenting all affixes of a word. Thus, all the prefixes and suffixes which are attached to the stem are separated. The morphological segmentation for the example mentioned earlier ومكتبته could be: و+مكتب+ت+ه As one may notice here, the suffix ة (feminine marker) is separated from the word. The parse tree of this word after full segmentation is as follows:



- **Arabic Treebank (ATB) segmentation :** this is a light segmentation adopted to build parse trees in the Arabic TreeBank (ATB) corpus. This type of segmentation considers splitting the word into affixes if and only if it projects an independent phrasal constituent in the parse tree. As an example, in the word ومكتبته mentioned earlier, the phrasal independent constituents are: conjunction و noun and the head of a Noun Phrase(NP) مكتبة and a pronoun ه (his). This would lead to the following parse tree:



• Developing a tagset for automated POS tagging in Arabic

People worked on it : SHIHADEH ALQRAINY and ALADDIN AYESH

Link:
https://www.researchgate.net/publication/262372038_Developing_a_tagset_for_automated_POS_tagging_in_Arabic

The paper explains the approach of developing a tagset for the Arabic language that depends on the Arabic system of inflectional morphology called الاعراب.

the main focus was on how the grammarians analyses all Arabic words into three main part-of-speech (verb, noun , particle) these are further sub-categorized into more detailed part-of-speech which collectively cover the whole Arabic language.

This way of developing the tagset will open up a lot of possibilities which they provided a sample of it in the papers.

3-Tagger approach

We used the hybrid approach which uses two approaches (rule based, patterns), the rule based will check first on the word and see if any of the rules work on the word, and if not then the pattern based will check for any patterns that match the word pattern.

1-Rule Based approach: this was our main approach because its reliable and more efficient than the patterns approach, there are some sets of rules that can define what is the tag for this word and what is the tag for the next word, and these rules works in most cases on the Arabic language.

2-pattern approach: we used it as a complimentary approach to the Rule Based, to reach higher accuracy, we couldn't put it as our main approach because there is a high number of patterns in the Arabic language, for instance the word كتب have its own pattern then if you want it to be in the present tense it will be يكتب which is another pattern , then you can give it a feminine mark تكتب , there is a high number of patterns just for one word in the Arabic language, and that's what makes it hard to relay only on this approach.

4-Our Tag Set

Rule based: the following rules are implemented with the same order as its listed

- Conjunctions that are followed by a noun like: ليت, لعل, من, الى , على , في , حيث , ...etc.
- Conjunctions that are followed by a verb like: ان , لن, كي ,حتى,مهما, ... etc.
- Conjunctions that are followed by a Conjunction like: اياك, اياكم , اياهم, ...etc.
- Kan verbs followed by a noun like :كان,ليس,امسى,ظل,...etc.
- Words with prefixes that tag it as a noun like:ال,لل,فال,كال,بال .
- Words with prefixes that tag it as a verb like:سي , سن ,ست ,سأ .
- Any verb will be followed by a noun like :قال محمد , ذهب الرجل, ...etc .
- The particle و the tag for the word before it is the same as the word after it like: محمد و علي, يأكل و يشرب, ...etc .

Patterns:

Patterns that will tag the word as a verb: يتفاعلان,افعوعل,افتعل, يتفعل, يتفاعل, استفعل, افعال , يستفعل, تفوعل, يفعلون, تمفعل , يستفعلان, تتفاعلون,تستفعلون,تفعلان, يفعلان,تفعلون,يتفعلون, يتفاعلون,يستفعلون,يتفعلان.

noun patterns: فعلة,فعيل,فعلى,مفعلة,مفاعلة, مفعال, مفعل, مفعول .

5-Conclusion

In this paper, we explained how our part-of-speech tagging going to work what rules we used and other techniques and approaches of how to analyses the Arabic language and how to create part-of speech tagger that can reach high accuracy and reliability, with the use of new technologies like deep learning and neural networks.

The part-of-speech tagging in the Arabic language is open to changes and enhancements, but it needs a lot of time and research, because the Arabic language is one of the few languages that have a high level of complexity. For instance one word in the Arabic languages can give a lot of meanings based on where it’s placed in the sentence and each root word(stem) can derive a lot of words from it that gives you a different meaning for it. And then you can add prefixes and suffixes to the word to give it a specific characteristic like plural or feminine mark.