*Journal of Autonomous Intelligence* (2018) Volume 1 Issue 2.
doi: 10.32629/jai.v1i2.30

## Original Research Article

# New Combined Method to Improve Arabic POS Tagging

***Mohamed Labidi***

*LaTICE laboratory, Tunisia*

## ABSTRACT

One of the important tasks in Natural language processing is the part of speech tagging. For the Arabic language we have a lot of works but their performances do not rise to the required level, due to the complexity of the task and the Arabic language characteristics. In this work we study a combination between two different approaches for Arabic POS-Taggers. The first one is a maximum entropy-based one, and the second is a statistical/rule-based one. Fur-thermore, we add a knowledge-based method to annotate Arabic particles. Our idea improves the accuracy rate. We passed from almost 85% to almost 90% using our combined method, which seem promoter.

***Keyword:*** *POS-Tagger; Natural language processing; Arabic language*

**ARTICLE INFO**
Received: Nov 28, 2018
Accepted: Dec 31, 2018
Available online: Jan 8, 2019

**CITATION**
Mohamed Labidi. New Combined Method to Improve Arabic POS Tagging. *Journal of Autonomous Intelligence* 2018; 1(2): 23-28. doi: 10.32629/jai.v1i2.30



## 1. Introduction

Any part-of-speech tagger aim to assign each word in a text a word class, or we can call it part of speech or tag. For example in the Arabic language a word can be a verb, a noun or a particle. So, an Arabic POS-Tagger should assign to each word in a given text a word that indicates its class (verb, noun or particle). The added tags to any text fills several purposes and is a very important resource for finding certain patterns in a lan-guage, analyzing word frequencies, syntactical structures and other analysis.

Below we present some of the most known Arabic taggers that have been funded during the background research.

The first important work to present is the Stanford Part-Of-Speech tagger. This tagger was developed at Stanford University and is described in (Toutanova and Manning, 2000). It was originally developed for Eng-lish. Other languages have been supported in the improved version described in (Toutanova et al., 2003). The Stanford POS-tagger is based on the maximum-entropy model.

The second one is the Brill tagger. In this tagger combinations of rule-based approach with a general ma-chine-learning approach. The idea behind is to initially let the text pass through an annotator, in part-of-speech context this might be assigning each word its most likely tag. Then the text is compared to the gold standard (the "best" tagging"), in order to create transformations that can be applied to improve the initial text as much as possible. A transformation is described in (Brill, 1995) it consists of two parts.

The third important one is the Khoja Arabic Part-Of-Speech Tagger This tagger also combines a statistical and rule-based approach as it is proven to produce the best results (Khoja, 2001). The tagger uses a tag set of 131 tags, which is derived from traditional Arabic grammatical theory. Four different corpora were used for testing, the largest consisting of roughly 59,000 words was used to train the tagger and to build various lexicons which is used when tagging the test set. One version of a lexicon lists each word together with all tags it has received in the corpus. This was used in the initial stage of the tagging by looking

up each word in the lexicon and gives all possible tags as specified in the lexicon. Stemming was used for words which were not found in the lexicon. For disambiguating tags, the Viterbi algorithm was used together with lexical- and contextual probabilities. In all, the tagger achieved an accuracy of around 90%. For further detail, see (Kho-ja, 2001).

The fourth one is the Tree tagger. This is a language-independent tagger by (Schmid, 1994) and is based on decision trees. The tagger has been successfully trained and tagged many European languages and it's adaptable to other languages if a lexicon and a manually tagged training corpus are available. The tagger comes with code for different architectures such as PC, Solaris and Macintosh. This gives an indication that it's likely written in C/C++. Wrappers for Java and Python are available by the community.

(Algrainy et al.,2008) described a novel idea in their work. They report an accuracy of 91%. They used a pattern-based approach to tag words, using a small manually annotated lexicon. Also another new idea was described in (Yousif and Sembok, 2008), the authors used the Support Vector Machines (SVM) approach and a corpus of 177 tagged words. They report a staggering accuracy of 99.99%.

In (Hatim Ibrahim, 2014) the author proposed a new approach to Arabic Part-of-Speech tagging based on augmented stateful sliding-window (SWPoST). His system assigns the part of speech to the word based on the information provided by a variable width window of words around it, called semantic window.

Also the part of speech tagging task can be made by another tool called NOOJ. It is based on linguistic approach. And it work well with English language but not with Arabic language because of the lack of resources.

Al-Khalil is tool created by (Boudlal et al, 2010). It is based on a database of annotated words. Almost 50.000 Arabic words.

Another works can found in (Dat et al, 2016). Ones are based on the hidden Markov models and others on statistical approaches. But the majority does not exist on the internet as a free tool or they are damaged.

From all the previous cited part of speech taggers, just few of them are available on the Internet, and can be used. Each one of the founded tools has its advantages and disadvantages. So we proposed in this paper a new method to combine these tools and make a new more powerful combined tool. It is not only a combination of tools, but it is also a combination of approaches where each tool based on an approach. Also in this paper we propose a new extension to ameliorate the Arabic POS-Tagger combined. The proposed method is described in the next section.

## 2. Proposed Method

This section contains a description of our combined POS tagging method. The following figure describes our idea.

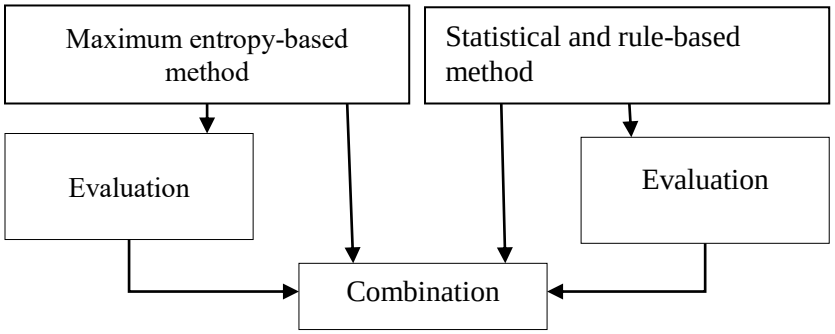


**Figure 1:** Combination of Arabic POs-tagging methods.

Each rectangle in Figure 1 presents a step in our combined approach. Our goal is to find the best way to combine a maximum entropy-based method and a statistical and rule-based method. To do this we should pass first of all by a step of evaluation, where each one of the two method evaluated by our test corpus. After that we make our combination method based on this evaluation. The following sections describe this combination.

## 2.1 Experimental setup

To evaluate the methods and the proposed combination, we used our textual corpus which is manually annotated. The following table is a presentation of our corpus of test. The Arabic language composed from three principal types of words, which are verbs, nouns and particles. We constructed a test set composed of 70 files that cover different themes (Sport, Politics, etc.). The created corpus contains 19,322 words. Each word in the corpus was manually annotated.

**Table 1.** Information about corpus of test

| Number of words | Number of nouns | Number of verbs | Number of particles | Punctuation |
|---|---|---|---|---|
| 19,322 | 10,300 | 5,701 | 1620 | 1701 |

## 2.2 Methods evaluation

Before describing our proposed combination, we should evaluate the maximum-entropy method and the statistical and rule-based method to compare them.

**Table 2.** Tools performances

| Method | Nouns annotation accuracy | Verbs annotation accuracy |
|---|---|---|
| Maximum entropy-based (Stanford tool) | 81% | 79% |
| Statistical and rule-based (Khoja tool) | 74% | 85% |

As presented in Table 2. The maximum entropy based method report an accuracy of 81% for the nouns annotation and 79% for the verbs annotation. The statistical and rule based method report an accuracy of 74% and 85% for the same tasks.

In the rest of the work we attempt to combine these two methods in order to increase the annotation rate and improve the Arabic POS-tagging.

## 2.3 POS-Taggers combination

Figure 3 describes the combination of the two methods selected in the previous step.

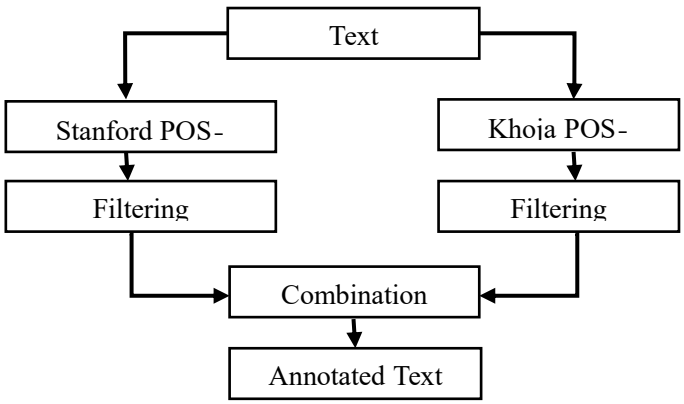


**Figure. 2.** Combination approach

For the annotation of the input text we follow the following algorithm:

*Begin*

*1) Annotate the input text by the two methods :*

*a. The maximum entropy method*

*b. The statistical and rule-based method*

*2) For the nouns in the text, only keep the annotation of the maximum entropy method because it is much better in this task (Table 2).*

*3) For the verbs in the text, only keep the annotation of the statistical and rule-based method because it is much better in this task (table 2).*

*End.*

Two problems appear during the execution of the previous algorithm:

1- Some nouns are ignored by the maximum entropy method.

2- Some verbs are ignored by the statistical and rule-based method.

To deal with this problem we modified the previous algorithm by the following one:

*Begin*

*1) Annotate the input text by the two methods :*

*a. The maximum entropy method*

*b. The statistical and rule-based method*

*2) For the nouns in the text, only keep the annotation of the maximum entropy method because it is much better in this task (Table 2). If the maximum entropy method ignored a noun, then keep the annotation of the statistical and rule-based method.*

*3) For the verbs in the text, only keep the annotation of the statistical and rule-based method because it is much better in this task (table 2). If the statistical and rule-based method ignores a verb, then keep the annotation of the maximum entropy method.*

*End.*

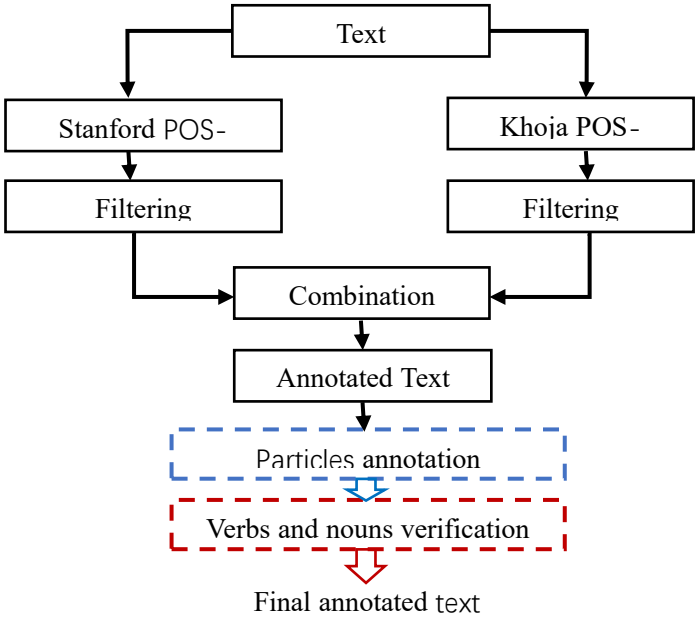


**Figure. 3.** The added Extensions

To ameliorate the annotation of the combined methods we added a new rules-based extension to ameliorate the verbs and the nouns annotation. Also we added a knowledge-based extension to annotate particles in the Arabic texts. The following figure describes the added extensionns.

The first extension that we added is the “Particles annotation”. This step is knowledge-based approach. We collected a dictionary that contains all the particles that can appear in an Arabic text. We found that in the Arabic language we have 80 particles. We used this dictionary to annotate the particles in the analyzed text. To annotate any word, we look for it in the dictionary if it exist then it will be annotated as particle. The following algorithm describes this process.

*Begin*
*For each word “W” in the transcription*
*If “W” exist in “Particles_Dictionary” then*
*Annotate “W” as particle.*
*End if*
*End for*
*End.*

The second extension that we added aims to ameliorate the proposed combination. It is a rule-based approach called in Figure 4 “Verbs and nouns verification”. In the Arabic language each particle has an effect on the words attached to it in the text. Some particles appear only with verbs and other ones appear only with nouns, and some appear with both of them. The next examples describe this phenomenon.

So we decided to use this particularity and to create a set of linguistic rules to verify the annotated text. The next rule is an example of the created rules.

**Table 3.** Example of rules

| | | | | |
|---|---|---|---|---|
| لم | + | word | ➔ | Word =noun |

In the previous rule, if we found the particle “لم “ with any word then this word is a noun and it will be annotated as noun because this particle appears only before nouns.

We create 50 rules based on this particularity of the Arabic language.

**Table 4.** Examples of Arabic particles effects

| | | | |
|---|---|---|---|
| Example of particles that appear only with nouns | Particle | IPA | لَمْ يَكُنْ لَهُ كُفُوًا أَحَدٌ<br>lam yakun lah kufuana 'ahad<br>Nor is there to Him any equivalent |
| | لمّا | Lmma: | |
| | لم | lm | |
| | ألم | ʔgm | |
| Example of particles that appear only with verbs | Particle | IPA | عَلَى الْفُلْكِ تُحْمَلُونَ<br>ealaa alfulk tuhmalun<br>And upon them and on ships you are carried |
| | بـ | b | |
| | في | faj | |
| | على | ʕla: | |

## 3. Tests and results

In this section we evaluate our new proposed method and the proposed extensions. We obtained the following results.

**Table 5.** Results of test

| Method | Nouns annotation accuracy | Verbs annotation accuracy |
|---|---|---|
| Maximum entropy-based (Stanford tool) | 81% | 79% |
| Statistical and rule-based (Khoja tool) | 74% | 85% |
| **Our method** | 90% | 92% |

Table 5 describes the obtained results during the tests. Our proposed method gives us a better accuracy. We consider these results as encouraging and promoted.

## 4. Conclusion

In this paper we proposed a new method to ameliorate the existent Arabic POS-Taggers. We made a new combination between two approaches (maximum entropy, statistical and rule-based). Furthermore, we added two new extensions. The first extension is an acknowledge-based one. And the second is rule-based one. We obtained a better accuracy using our new method and we hope to ameliorate it in the future.

## 5. References


Ababou N, Mazroui A (2016) A hybrid Arabic POS tagging for simple and compound morphosyntactic tags. International Journal of Speech Technology 19:289–302.

Algrainy S, AlSerhan H M, Ayesh A (2008) Pattern-based algorithm for part-of-speech tagging. In International Conference on Computer Engineering and Systems, ICCES, pages 119-124.

Ann Bies. http://www.ircs.upenn.edu/arabic/Jan03release/arabic-POStags-collapseto-PennPOStags.txt

Berger A, Della Pietra S, Della Pietra V J (1996) A maximum entropy approach to natural language processing. Computational Lingustics, 22(1):39-71.

Brill E (1995) Transformation-based error-driven learning and natural language processing: A case study in part-of-speech tagging. Computational Lingustics, 21(4):543-566.

Jurafsky D, Martin J (2009) Speech and Language Processing. Pearson Education.

Khoja S (2001) Apt: Arabic part-of-speech tagger. In Proceedings of the Student Workshop at the Second Meeting of the North American Chapter of the Association for Computational Linguistics.

Khoja S, Garside R, Knowles G (2001) A tagset for the morphosyntactic tagging of Arabic. Corpus Linguistics.

Rabiie H S (2001) Arabic Language Analysis Toolkit. http://www.nada.kth.se/utbildning/grukth/exjobb/rapportlistor/2011/rapporter11/shouhani_rabiee_hajder_11

133.pdf

Schmid H (1994) Probabilistic part-of-speech tagging using decisions trees. In International Conference on New Methods in Language Processing.

Toutanova K, Manning CD (2000) Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. In Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC-2000), pages 63-70.

Toutanova K, Klein D, Manning C, Singer Y (2003) Feature-rich part-of-speech tagging with cyclic dependency network. In Proceedings of HLT-NAACL, pages 252-259.

Yousif J H, Sembok T (2008) Arabic part-of-speech tagger based support vector machines. In International Symposium on Information Technology.

Boudlal A, Lakhouaja A, Mazroui A, Meziane A, Bebah M (2010) Alkhalil morpho sys1: A morphosyntactic analysis system for arabic texts. In International Arab conference on information technology (pp. 1-6). Benghazi Libya.

Dat Q N, Dai Quoc N, Dang D P, Son B (2016) A Robust Transformation-Based Learning Approach Using Ripple Down Rules for Part-Of-Speech Tagging. AI Communications (AICom), vol. 29, no. 3, pp. 409-422.