# Rule Based Part of Speech Tagger for Arabic Question Answering System

**Samah Ali Al-azani and C. Namrata Mahender**

**Abstract** Part of Speech (POS) Arabic wording is difficult to read in detail and its functionality affects many programs and activities in the Natural Language Processing (NLP) area. POS tagging is a process to assign POS such as a verb, adjective, adverb, noun in each word for any sentence. Farasa is an active and reliable text processing toolkit for Arabic documents. It is an assortment of Java libraries and CLIs for MSA.2. These incorporate a separate tool for Arabic text Diacritizer, segmentation/tokenization module, POS tagger, Named Entity Recognition (NER), and parsing. One of the limitations over the Farasa affects the post-processing results due to the presence of inappropriate tags. For our application on question answering system(QAS) correct POS, tagging is essential for better accuracy. The POS tagger is developed using a rule-based approach which is based on domain-specific. The corpus (database) on which the rule-based POS tagger is built is centered on the core subject of the Arabic language 4th standard textbook of Arabic Medium state board of Yemen. During the development of QAS, the POS tagger is a very essential stage in which the answers for the framed questions are obtained from the paragraphs of a given lesson. The present article provides insights into the complete process of linguistic rule-based POS tagger development for QAS. Sentence segmentation, word tokenization, to stemmer development which becomes an important part of proper morphological analysis is explained. As a result, the morphological analyzer is the input to the rule-based POS tagger. Ultimately, in this article, a comparison of marking based on our POS rule with Farasa is presented and for QAS, our rule-based POS tagger gave better results than Farasa.

**Keywords** Arabic language · Natural language processing (NLP) · Part of speech (POS) tagging · Morphological analyzer · Stemmer · Tokenization

S. A. Al-azani (✉) · C. Namrata Mahender
Department C.S. and I.T, Dr. Babasaheb Ambedkar, Marathawada University, Aurangabad, Maharashtra, India
e-mail: alazani183@gmail.com

C. Namrata Mahender
e-mail: nam.mah@gmail.com

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2021

V. Bindhu et al. (eds.), *International Conference on Communication, Computing and Electronics Systems*, Lecture Notes in Electrical Engineering 733,
https://doi.org/10.1007/978-981-33-4909-4_58

# 1 Introduction

Natural language processing (NLP) is a category of artificial intelligence that assists computers to recognize, translate, and administer human language. It is utilized in numerous disciplines and incorporates computational linguistics, computer science, and so on. It also bridges the gap between computer understanding and human communication [1]. This methodology recognizes natural language texts and linguistic procedures namely, part of speech (POS) tagger, lexicon, and tokenization. These are applied to recreate inquiries into the right inquiry that separates the significant answers from a structured database [2]. The inquiries dealt with by this methodology are of Factoid type and have a profound semantic understanding [3].

## *1.1 Arabic Natural Language Processing*

The Arabic language is a combination of many variations among different similarities that have a specific event, such as the standard official text of media and education throughout the Arabian World [4]. It is one of the top languages in the world. Its phenomenal script, distinguished style, and strong vocabulary confer a unique character and characteristic to the language. Arabic is the largest member of the Semitic language family. Nowadays, it is an official language in more than 20 countries and over 300 million native speakers [5]. Concerning other Semitic dialects, Arabic morphology was set up around the abstract idea of the root, three consonants articulating to significance, regardless of whether exact or ambiguous. The conventional derivational morphology is dependent on the root-and-pattern model concentrating on this abstract consonantal root. A pattern is an intermittent affix (or transfix) composed of vowels and non-revolutionary consonants embedded around spaces for the root consonants. To each pattern, conventional syntax relates a morphological classification or potentially inflectional highlights. These formalizations are applied by traditional grammar to depict both inflectional and derivational morphology. Moreover, the quantity of ‘voweled stem canonic patterns’ for action words and nouns is almost 10,000 [6]. Some problems in Arabic texts include considerable translations and translated labeled entity, its satire is usually contradicted texts on Arabic. Despite the fact that the methodology demonstrated to create high exactness, it has a few flaws. It requires building a tremendous corpus (dataset) and marking it manually by human experts. The system of manual commentary can be extremely troublesome in any event, for native speakers because of criticism and social references. It can also be costly and tedious. A constraint is that NLP tools designed for Western dialects are not effectively versatile to Arabic due to the specific highlights of the Arabic language [7].

# 2 Related Work

Many different researchers have focused on part of speech tagging in many languages like Hindi, English, Arabic, Chinese, Marathi, and others. Here some important related work in part of speech tagging techniques is discussed.

Singa et.al. (2012) has advanced part of speech tagger on based rule approach with a rate of accuracy 50, 77, 85% on lexicon data words (50-100-1000) in Manipuri language (Manipuri is mightily Speaker in Manipur, Bangladesh, Tripura, and Myanmar, Assam [8].

Zelalem (2013) has used a hybrid approach of HMM and rule-based tagger, on a collected dataset of 354 sentences with accuracy of 77.19, 61.88, and 80.47% in Kafi- **Noonoo** language (language in southwestern Ethiopia) [9].

Deepali et al. (2018) has designed the POS tagging base rule in the Marathi language, on a collection of 1364 words with an accuracy rate was 100% [10].

Aliwy et al. (2018) suggested a new approach using HMM and n-grams taggers for tagging Arabic words in a long sentence to collect 1000 datasets as documents and 526,321 as a separate token, this system gives an accuracy rate of 0.888, 0.925, and 0.957 [11].

Barud et al. (2019) developed parts of speech tagger for Awngi language using Hidden Markov Model (HMM), in a dataset 94,000 sentences were collected, total word of 188,760 and attained accuracy rate of 93.64 and 94.77% [12].

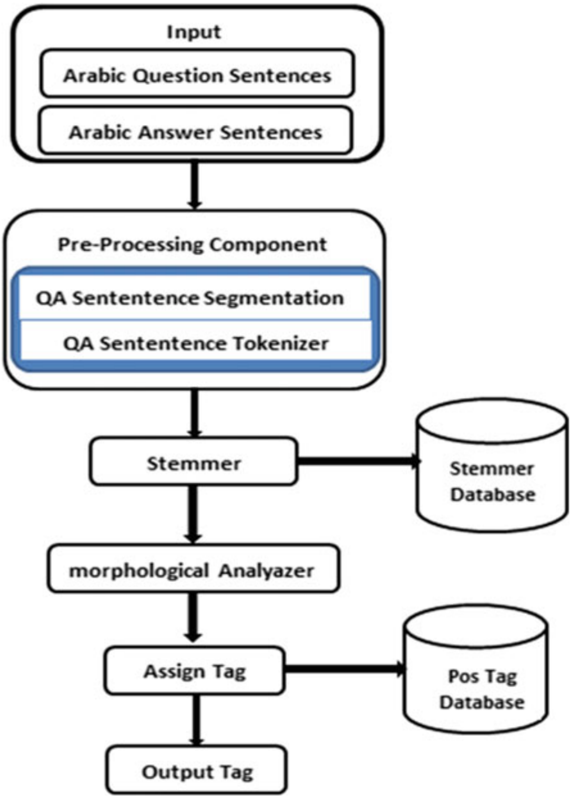
Hagos, 2020 designed a Ge'ez (Ethiopian language) POS tagging using a Hybrid approach, the data set collected was 15,154 words, 1,305 sentences, and the result accuracy rate was 77.87, 82.23, and 94.32% [13].A detailed description of building a POS tagger is discussed in further sections.

# 3 Proposed System

## *3.1 Data Collection*

The proposed system considers 40 questions and 40 answers that are taken from five different lessons of the 4th standard textbook of the Arabic language as the raw input. The data is pre-processed, and a well-designed stemmer is developed for supporting the morphological analyzer. An output of a morphological analyzer is used as an input to the part of speech tagging. A linguistic rule-based POS tagger is developed, and ultimately, the tag data is provided by the system. This tagged data will be used in post-processing for better performance of the QA system. The detailed block diagram of the system proposed is shown in Fig. 1, and further stages have been discussed in detail with example.

**Fig. 1** Proposed system



## 3.2 *Pre-processing Component*

### 3.2.1 Questions and Answers Sentence Segmentation

In the Arabic language, the structure is dissimilar for different sentence types. The proposed system consists of two segments namely, question and answer. The former is the question part, and the latter is the answer part. The input data is taken from the 4th standard textbook of Arabic language. The questions are framed using the textbook lesson and the number of word used are confined. The answers for the question are stored in the answer file and the answers are based on the lessons.

### 3.2.2 Questions and Answers Sentence Tokenization

Tokenization is the method of tokenizing or parting a string, text into a listing of tokens. One can consider token parts like a word is a token in a sentence, and a sentence is a token in a passage. This process of separating tokens is the input text.

Each word is separated from the sentence considering white space or symbols as one token and treat each word individually. Then POS needs to split the input text into tokens tagging. The tokenization of the question sentence is shown below:

' لماذا يعد البن اليمني من اجود انواع البن في العالم ؟ '

and word tokenization has split answer sentence into words as follows:

' يعد' ' البن' ' اليمني' ' من' ' اجود' ' انواع' ' البن' ' في' ' العالم' ' لما' ' يمتاز' ' بة
' ' من' ' نكهة' ' طيبة ' '.'

## 3.3 Stemming

Stemming is the operation of decreasing a word to its word stem that contains affixes to suffixes and prefixes or the roots of words known as the lemma. A rule-based steamer uses specific pre-defined rules according to language to mark another type of word used in its base. These language connected rules are created manually by language practitioners. Rule-based stemming methods are divided into three categories such as morphological, table lookup, and affix stripping. Arabic stemmer is a very different and difficult structure than other languages. Stemming is very necessary for natural language understanding and natural language processing. Arabic word language structure is a grammatical rule on the root, and pattern scheme, its deliberate as a root based language with more than 10,000 roots. Arabic words are commonly founded on three-dimensional roots: three consonants, which characterize the hidden significance of the word. Diverse long and short vowels, prefixes, and postfixes are added to that root to make the ultimate desired inflection of sense. These modifications follow designs that reflect across roots. Stemming is the only source to extract the root in the Arabic language. For example, the Arabic word ' المعلمات 'contains the following component in Tables 1 and 2.

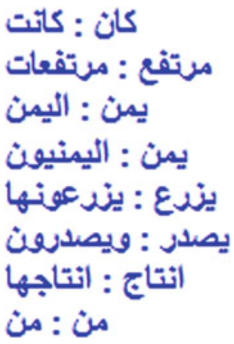
Figure 2 depicts an example for code execution and the flow of stemming in the proposed model (Figs. 3 and 4).

## 3.4 Morphological Analyzer

The morphological analyzer of Arabic words is a procedure for each word of the input text to choose its root and pattern. The outcomes of the morphological analyzer can be used for further analysis[14]. The morphological analysis goals are to train the internal structure of a word. Words after molding are analyzed to check if they are sorted or not. When a stem word is produced, then the word root is formed

**Table 1** Example of Arabic Affix

| Word | root | prefix | Suffix | Infix |
|---|---|---|---|---|
| المعلمات | علم | ال | ات | ا |



```
كان : كانت
مرتفع : مرتفعات
يمن : اليمن
يمن : اليمنيون
يزرع : يزرعونها
يصدر : ويصدرون
انتاج : انتاجها
من : من
```

**Fig. 2** Result for Arabic stemming from QA Sentences

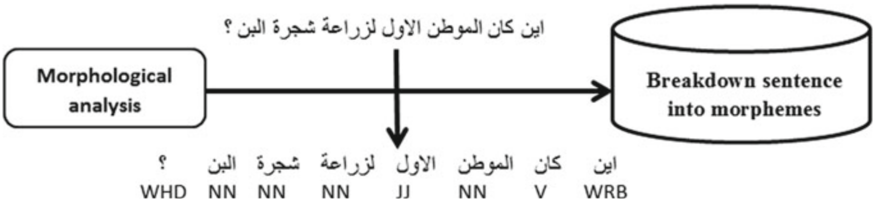
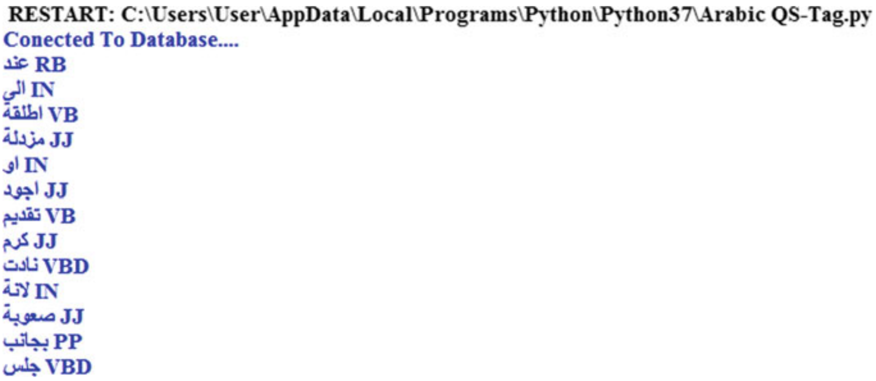


**Fig. 3** Role of the morphological analysis of Arabic words



```
RESTART: C:\Users\User\AppData\Local\Programs\Python\Python37\Arabic QS-Tag.py
Conected To Database....
عند RB
الي IN
اطلقة VB
مزدلة JJ
او IN
اجود JJ
تقديم VB
كرم JJ
نادت VBD
لانة IN
صعوبة JJ
بجانب PP
جلس VBD
```

**Fig. 4** Result of rule-based POS tagger for QA in the Arabic language

by combining substituted letters with the stem word. The morphological analyst is expected to produce root names for the given input document [2].

**Table 2** Arabic prefixes

| Words Example | Prefix |
|---|---|
| بالحديقة | ب |
| كدواء | ك |
| والاولاد | و |
| الزراعة | ال |

**Table 3** Rule-based POS tagging and Farasa POS tagging

| Rule-based POS tagging | Farasa POS tagging | Word | NO |
|---|---|---|---|
| RB | NOUN | عند | 1 |
| IN | NOUN | الى | 2 |
| VB | NOUN | اطلقة | 3 |
| JJ | NOUN | مزدلة | 4 |
| IN | NOUN | او | 5 |
| JJ | NOUN | اجود | 6 |
| VB | NOUN | تقديم | 7 |
| JJ | NOUN | كرم | 8 |
| VBD | NOUN | نادت | 9 |
| IN | NOUN | لانة | 10 |
| JJ | NOUN | صعوبة | 11 |
| PP | NOUN | بجانب | 12 |

## *3.5 Tag Generation*

Initially, Farasa is attempted to obtain the tagged word, but the results were not that much appreciable. Table 3 shows the incorrect tags using Farasa.

A linguistic rule-based POS tagger is chosen to be developed due to the limitations over the Farasa.

## *3.6 Rule Based POS Tagger*

The data is manually analyzed to find the context for each part of speech and based on the data a rule is developed.

### (a) **Algorithm for POS Tagging System**

**Step 1** Input the question and answer sentence segmentation.

**Table 4** Proposed system data

| Lessons | No. of (Q & A) | No. of words (Q & A) |
|---|---|---|
| L1 | 16 | 134 |
| L2 | 16 | 100 |
| L3 | 16 | 123 |
| L4 | 16 | 135 |
| L5 | 16 | 131 |
| Total Words | | 623 |

**Step 2** Start to tokenize the Q & A sentence into words.

**Step 3** Generate a stemmer for all words to obtain the original words by using a morphological analyzer.

**Step 4** Gather all the incorrect tags for all the words from the Farasa parser as a word by word and transfer to the database to produce the correct tag for all words.

**Step 5** Allocate a suitable tag to append to its word.

(b) **Result**

In this article, the data were collected from the Arabic language among one of the subjects from the 4th standard Arabic medium state board pattern in Yemen. Five different lessons are considered in this work and formed 40 questions and 40 answers respectively. Table 4 illustrates the data in lessons.

The formula to calculate the accuracy and the performance in Farasa parser tagging and rule-based POS tagging is applied:

**Accuracy = (No. of Correctly tagged/ Total No. Tagged in documents)*100.**

Subsequently, the accuracy result in Farasa parser was 92.77%. and the proposed system accuracy is 100%.

## 4 Conclusion and Future Scope

Question Answering System comes beneath the purview of natural language processing, thus natural language understanding is an essential task. For the present work, the corpus (database) for the QAS is build based on the Arabic language subject of the 4th standard Arabic medium state board pattern in Yemen. POS tagger is a very important component of any QA system as it impacts the accuracy of question answers generated for the system. In thisarticle, the major inputs to the POS tagger is based on sentence splitting, word tokenization, and stemming which serves as the basis for better performance of POS tagging. Linguistic POS tagger rule-based design and implementation are presented in detail concerning the QA system. The results are compared with Farasa and determined that our rule-based system yielded better results. In the future scope, this tagged data will be employed to measure the results of the QA system.

## References

1. Shaalan K, Siddiqui S, Alkhatib M, Monem AA Challenges in Arabic natural language processing , School of Informatics, University of Edinburgh1, UKFaculty of Computer and Information Sciences, Ain Shams University 4, Abbassia, 11566 Cairo, Egypt
2. Yao X (2014) Feature-driven question answering with natural language alignment , Johns Hopkins University, Doctor of Philosophy thesis 2014
3. AL-Taani A, Abu Al-Rub S (2009) A rule-based approach for tagging non-vocalized Arabic words. Int Arab J Inf Technol 6(3)
4. Habash NY (2010) Introduction to Arabic natural language processing, A Publication in the Morgan and Claypool Publishers series. ISBN: 9781598297966
5. https://arabicquick.com/an-introduction-to-the-arabic-language
6. Darwish (2002) Building a Shallow Arabic morphological analyzer in one day. In: Proceedings of the ACL workshop on computational approaches to semitic languages, Philadelphia, PA, pp 1–8
7. Ezzeld AM, Shaheen M (2012) Survery of Arabic question answering: challenges ,tasks , approaches , tools , and future trends. In: The 13th international Arab conference on Information technology ACT2012. ISSN.1812–0857
8. Raju Singha KH, Purkayastha BS, Singha KD (2012) Part of speech tagging in Manipuri: a rule-based approach. Int J Comput Appl 51(14): 0975–8887
9. Mekuria Z (2013) Design and development of part-of-speech tagger for Kafi-noonoo language, Master's thesis, Addis Ababa University, Addis Ababa
10. Deepali G, Naik Ramesh R, Namrata Mahender C (2018) Rule-based part-of-speech tagger for Marathi language, © 2018 IJSRST | vol 4, issue 5. Print ISSN: 2395–6011. Online ISSN: 2395–602X Themed Section: Science and Technology
11. Aliwy AH, Al_Raza DA (2018) Part of speech tagging in Arabic long sentence. Int J Eng Technol 7(3.27):125–128
12. Wubetu Barud Demi lie, Parts of Speech Tagger for Awngi Language. ISSN 2321 3361 © 2019 IJESC
13. Gebremedhin H, gebremeskel S (2020) Ge'ez POS tagger using hybrid approach. Int J Comput Sci Inf Technol Res 8(1):12–23
14. Awajan A (2018) A rule-based morphological analyzer of Arabic words. https://www.researchgate.net/publication/330006249