Proceedings of the New Trends in Information Technology (NTIT-2017).
The University of Jordan, Amman, Jordan. 25-27 April 2017.

PART-OF-SPEECH TAGGING FOR CLASSICAL AND MSA ARABIC TEXT USING NLTK

KhetamYassen, MajdiSawalha, Fawaz Al Zaghoul

Computer Information Systems Department, King Abdullah II School for Information Technology, The University of Jordan

E-mail: khetamyassen@yahoo.com, sawalha.majdi@ju.edu.jo, fawaz@ju.edu.jo

Abstract

This paper aims to investigate the problem of Part of Speech (POS) tagging for Arabic text. POS tagging is an essential text analytics application. It serves as a prerequisite for many other applications. POS tagging for Arabic text is still unsolved problem. Most POS tagging algorithms are designed for Modern Standard Arabic (MSA). This study investigates the applicability of these POS tagging models for Classical Arabic (CA) and MSA text. Our methodology depends on building POS tagging model using eight different algorithms(Unigram Tagger; Bigram Tagger; Trigram Tagger; N-gram Tagger; Brill Tagger; Affix Tagger; HMM Tagger and TnT Tagger provided by the Natural Language Toolkit (NLTK). We used the Qur'an text as a gold standard for modeling POS tagging for CA text. Different POS tagging algorithms were investigated and applied to analyze CA text (i.e. Qur'an text). Standard evaluation metrics were used to compare the performance of the different Machine Learning (ML) algorithms. The aim of this paper is to transfer knowledge from Quran text to MSA text to prove Quran text as a gold standard. The results of different POS taggers were compared, studied, and represented in this paper.

Keywords: Classical Arabic, Machine learning, Natural language Toolkit, Modern Standard Arabic, Part of Speech Tagging, NLTK.

1. INTRODUCTION

All languages in the world have a particular importance, because the language is a tool of the expression and imagination of human feelings and emotions, as in [14]. Lingual issues of a certain language contain social, commercial, cultural, political, educational, academic, media, sport and entertainment elements. These issues can never be separated from other issues, such as; thinking, communicating, building relationships, selling, buying, educating, teaching and entertaining , as in [4].Researchers estimate number of Arabic native speaker ranges between 280 and 400 million Arabic native speakers. Arabic is the fifth widely used language in the world, and one of the official languages of the United Nations (UN) since 1974, as in [2].

The Arabic language is a Semitic language originated in the Arabian Peninsula. The Modern Standard Arabic (MSA) is the language mainly used in the media (radio, TV, news bulletin, books, journals, newspapers, ads, etc...) and it is the language used in education and official correspondence. MSA dates back to the end of the eighteenth century, and it is the official language of 27 countries worldwide. The MSA is a descendant of the CA language (the language of the holy Qur'an) that was used in the 6th century. MSA and CA are different mainly in style and vocabulary. The Arabic language used in the Arab world is divided into two main versions: MSA and Colloquial (dialectal) Arabic. The MSA has no variants while Colloquial Arabic has many regional variants (dialects), as in [2]. A publicly available grammatically tagged corpus of Arabic still does not freely exist. Arabic is a morphologically rich language, as in[9]. Word carries not only inflections but also clitics, such as pronouns, conjunctions, and prepositions. This morphological complexity also has consequences for the POS annotation of Arabic. POS tags contain information about the morphology, i.e. they refer to segments rather than to whole words. Thus, the word وسيكتبونها (wsyktbwnhA, they will write it) is assigned the following POS tag, as in [11]:

و	w	CONJ +
س	s	FUTURE PARTICLE +
ي	y	IMPERFECT VERB PREFIX +
كتب	ktb	IMPERFECT VERB +
ون	wn	IMPERFECT VERB SUFFIX MASCULINE PLURAL3RD PERSON +
ها	hA	OBJECT PRONOUN FEMININE SINGULAR

This word form consists of a conjunction, a future particle, an inflectional prefix, the verb stem, an inflectional suffix, and a pronominal object. The boundaries between segments are depicted by + signs. As can be seen from this example, three of the segments (the conjunction, the future particle and the object pronoun) as well as the stem ktb, are syntactically independent although they are part of the orthographic form, i.e. they are clitics , as in [11].

POS tagging is process of assigning grammatical category labels to all words of a text according to their context. A tag is a grammatical category label such as noun, verb, adjective, adverb, etc, as in [1]. POS taggers are programs that are designed to assign the correct POS tag for each word in a text according their context. In general, POS taggers are essential technology for many text analytics applications. It serves as a prerequisite for many other applications such as: used in searching grammatical error detection in word processing; training neural networks for grammatical analysis of text; or training statistical language processing models; searching the web; syntactic analysis; constructing dictionaries; speech processing; word sense disambiguation; information extraction; document classification; sentiment analysis; document similarity; automatic summarization; grammatical error detection in word processing; etc.

Ambiguity of POS for Arabic text is still an unsolved problem, as in [9]. So we will design a solution to the problem of POS tagging for Arabic text using Quran corpus, and MSA corpus as training data. POS tagging algorithms implemented in the Natural Language Toolkit (NLTK) will be trained, tested and evaluated.

The Holy Qur'an is the religious text of Islam. The Qur'an is considered as an excellent gold standard text and essential for developing, modeling and evaluating Arabic NLP application. The Qur'an as corpus consists of 77,430 words. It is divided into 114 chapters which consist of 6,243 verses. Each chapter contains a sequence of numbered verses, as in [5][13].

The CA of the Qur'an has been relatively unexplored, despite the importance of the Qur'an to Islam worldwide which can be used by both scholars and learners, as in [13]. The text of the Qur'an is fully vowelized where short vowels appear on each letter of any word in the Qur'an. This information gives an advantage to when automatically annotating the Qur'an text compared to other forms of Arabic like MSA, as in [5]. The removal of short vowels from the Qur'an text increases POS ambiguity. Table 1 below shows examples of Qur'an words that show the POS ambiguity problem. Note that short vowels were removed from the examples listed in Table 1.

In this article, we will use the Boundary Annotated Quran Corpus (BAQ), as in [10] which was constructed as a dataset for ML application. The BAQ uses two tag sets which were designed based on Traditional Arabic grammar. The first tagset contains three tags {noun, verb and particle}.The second tag set uses ten major syntactic categories as they were defined in Quranic Arabic Corpus, as in [3]: {nouns, pronouns, nominals, adverbs, verbs, prepositions, lām prefixes , conjunctions, particles and disconnected letters}. The BAQ was used in this research as the main training dataset for ML algorithms to solve the problem of POS tagging for CA represented by the Qur'an text.

1.1 THE QUR'AN AS CORPUS

Table 1. Examples of words from the Qur'an that shows POS ambiguity problem

English translation	Part-of-Speech	Word	Word in Qur'anic Context
That is **the guidance** of Allah by which He guides whomever He wills of His servants.	Noun	هدى	ذلك هدى الله يهدي به من يشاء من عباده(6:88)
Those are the ones whom Allah **has guided**, so from their guidance take an example.	Verb	هدى	أولئك الذين هدى الله فبهداهم اقتده(6:90)
I did not make them witness to the**creation of** the heavens and the earth.	Noun	خلق	ما أشهدتهم خلق السماوات والأرض(51)
It is He who **created for** you all of that which is on the earth. Then He directed Himself to the heaven, [His being above all creation], and made them seven heavens, and He is Knowing of all things.	Verb	خلق	هو الذي خلق لكم ما في الأرض جميعا ثم استوى إلى السماء فسوهن سبع سماوات وهو بكل شيء عليم (29)
And **mentions** the name of his Lord and prays.	Verb	ذكر	وذكر اسم ربه فصلى (15)
But it is not except a **reminder** to the worlds.	Noun	ذكر	ما هو إلا ذكر للعالمين (52)

1.2 MODERN STANDARD ARABIC TEXT

The modern form of Arabic is called Modern Standard Arabic (MSA) and it is the form used by all Arabic-speaking countries in publications, the media and academic institutions. MSA is a simplified form of CA, and follows its grammar. The main differences between Classical and MSA are that MSA has a larger (more modern) vocabulary, and does not use some of the more complicated forms of grammar found in CA, as in [9].

Our article is organized as follows. First, we summarize what is known about Natural language toolkit (NLTK). Next we show how we can use taggers to emphasize learning from Quran text and MSA. Following this, we discuss the performance of POS tagging models when trained using the Quran dataset. Finally, the discussion and conclusion sections report on the results of building POS tagging models that uses the Qur'an as gold standard for training.

2. NATURAL LANGUAGE TOOLKIT (NLTK)

The Natural Language Toolkit (NLTK) issued in 2001 at the University of Pennsylvania in the Department of Computer and Information Science as part of a Computational Linguistics course. NLTK supports Accessing Corpora, String processing, collocation discovery, Part-of-speech tagging, classification, chunking, Parsing, semantic interpretation, evaluation metrics, probability and estimation … etc. NLTK is suitable for linguists, engineers, students, educators, researchers, and industry users alike. NLTK is available for Windows, Mac OS X, and Linux. NLTK is a free, open source, community-driven project, as in[3].

We have used NLTK's modules for building POS tagging systems for Arabic text. In this research we used nltk.tag and evaluation metrics from NLTK.

Supervised learning of taggers from POS-annotated training text is a well-studied task with several methods achieving near-human tagging accuracy, mean each word has tag as tuple (Word, Tag), as in [10].

We have used in this article nltk.tag and evaluation metrics with supervised to POS. In additional we have choice python program to deal with this model. Nltk.tag is handling with methodology.

3. NLTK'S PART OF SPEECH TAGGING ALGORITHMS

NLTK is a leading platform for building Python programs to work with human language data through the fundamentals of writing Python programs, working with corpora, categorizing text, analyzing linguistic structure, and more[1].

Researchers became interested in developing POS tagging systems for many languages and including Arabic. In this research, we will adopt POS tagging algorithms implemented in the NLTK for POS tagging Arabic text. Different NLTK POS tagging algorithms will be trained using the Boundary Annotated Qur'an (BAQ) corpus, evaluated and test using samples of both the Qur'an and MSA text. The following POS tagging algorithms were used in this research:

3.1 THE DEFAULT TAGGER

Default Tagger is Part of Speech Tagger Model that assigns the same POS tag for every token in a text. The default tagger takes a single argument (*i.e.* the tag that you want to apply). This tagger is inherited from the library nltk.tag, as in [8]. The below code segment shows the steps for creating a default tagger.

```
from nltk  import DefaultTagger
nltk.tag.sequential.DefaultTagger(tag)
```

3.2 THE UNIGRAM TAGGER

A unigram tagger model estimates the most likely tag for each word in a training corpus, and then uses that information to assign tags to new tokens, as in [8]. The below code segment shows the steps for creating a unigram tagger.

```
from nltk import UnigramTagger
nltk.tag.sequential.UnigramTagger (train = None, backoff=None, cutoff =0)
```

3.3 THE BIGRAM TAGGER

Bigram Tagger uses the previous tag as a part of context. Bigram tagger looks at two items (the previous tag and word), as in [8]. Bigrams are used in one of the most successful language models for speech recognition[2]They are a special case of N-grams where n = 2.

The below code segment shows the steps for creating a bigram tagger.

```
from nltk import BigramTagger

nltk.tag.sequential.BigramTagger(train=None,
model=None, backoff=None, cutoff=0,verbose=False)
```

3.4 THE TRIGRAM TAGGER

A Trigram Tagger uses the previous two tags which means it looks for three items (i.e. the previous two tags and the current word). A tagger that chooses a token's tag based on

[1]http://www.nltk.org/

[2]https://en.wikipedia.org/wiki/Bigram

its word string and on the preceding two words' tags[3] . The below code segment shows the steps for creating a trigram tagger.

```
from nltk  import TrigramTagger

nltk.tag.sequential.TrigramTagger (train=None,
model=None, backoff=None, cutoff=0, verbose=False)
```

3.5 THE N-GRAM TAGGER

N-gram Tagger is a subsequence of "n" items which means the tagger model looks for previous n-1 tags and words, as in [15]. The below code segment shows the steps for creating an n-gram tagger.

```
from nltk  import NgramTagger

nltk.tag.sequential.NgramTagger(n, train=None,
model=None, backoff=None, cutoff=0, verbose=False)
```

3.6 THE TNT TAGGER

TnT, the short form of "Trigrams'n'Tags", is a very efficient statistical POS tagger that is trainable on different languages and virtually any tagset, as in [12]. The below code segment shows the steps for creating a TnT tagger.

```
from nltk import tnt

nltk.tag.tnt.tnt.TnT(unk=None, Trained=False, N=1000,
C=False)
```

3.7 THE BRILL TAGGER

Brill's transformational rule-based tagger called Brill Tagger, uses an initial tagger to assign an initial tag sequence to a text which the tag is assigned to each word and changed using a set of predefined rules. The below code segment shows the steps for creating a Brill Tagger, as in [7].

```
from nltk import brill

nltk.tag.brill.brill_trainer.BrillTaggerTrainer (initial_tagger,
templates, trace=0, deterministic=None, ruleformat='str')
```

3.8 AFFIX TAGGER

Affix Tagger is a tagger for sequential sentences. Affix tagger treats trailing substring of its word string as an affix[4]. The below code segment shows the steps for creating an affix tagger.

```
nltk.tag.sequential.AffixTagger (train=None, model=None,
affix_length=-3, min_stem_length=2, backoff=None,
cutoff=0, verbose=False)
```

[3]Comment, Tag Freq Example. "4. Categorizing and Tagging Words."
[4] http://www.nltk.org/api/nltk.tag.html
[5]http://www.nltk.org/_modules/nltk/tag/hmm.html

3.9 HIDDEN MARKOV MODEL

A hidden Markov model (HMM) is constitute a family of versatile statistical models that have proven useful in many applications, as in [6]. HMM is a statistical Markov model in which the system being modeled is assumed to be a Markov process with unobserved (hidden) states. The mathematics behind the HMM were developed by L. E. Baum and coworkers. HMM largely used to assign the correct label sequence to sequential data or assess the probability of a given label and data sequence[5]. The below code segment shows the steps for creating a HMM tagger.

```
nltk.tag.hmm.HiddenMarkovModelTagger(symbols,  states,
transitions, outputs, priors, transform=<function _identity>)
```

4. METHODOLOGY FOR POS TAGGING OF ARABIC TEXT

This study aims to build a statistical POS tagging model for CA and MSA text. Our methodology depends on building POS tagging model using eight different algorithms provided by the Natural Language Toolkit (NLTK), as in [3].

These algorithms are Unigram Tagger; Bigram Tagger; Trigram Tagger; N-gram Taggers; TnT Tagger; Affixes Tagger; Brill Tagger and HMM Tagger (see Section 3).

The first POS tagging model will use the BAQ corpus, as in [10]. The BAQ corpus is a machine readable dataset designed for modeling ML algorithms (See Section 1.1). Figure 1 shows our methodology for building POS tagging models for CA and MSA text. The selected NLTK's POS tagging algorithms will be trained using the BAQ Corpus. These POS tagging models will be tested and evaluated using texts samples from the Qur'an representing CA a sample representing MSA.

5. EVALUATION OF PART OF SPEECH TAGGING ALGORITHM FOR ARABIC TEXT

5.1 EVALUATION USING QURAN CORPUS

POS tagging models for Arabic text have been evaluated in 10-fold cross validation using the Quran text (BAQ). BAQ Corpus contains 2 types of tagsets. The tagset consists of 3POS tags (noun, verb, particle) which represents the three main POS categories of traditional Arabic grammar. The second tagset consists of more detailed tags where the main POS categories were expanded to 10POS tags that include pronouns, nomionals and adverbs... etc (See Section 1.1).

The POS tagging models were trained and tested using the BAQ Corpus where POS tags and the words were the essential features for training. Two types of POS models were built. First, POS tagging models using the BAQ Corpus with 3POS tags. Second the type used 10POS tags and words are features for building these POS tagging models for Arabic.

The Evaluation of these models used 10-fold cross validation using BAQ Corpus. Each fold of train and test uses 90% of the BAQ Corpus for training and 10% for testing. 90% of the BAQ were used for training were formatted in sentences where each sentence is a tuple of word and tag. Testing sentences representing 10% of the BAQ Corpus were formatted as lists of words. The results of each experiment were compared against gold standards which were generated for each test sample from the BAQ Corpus. The gold standards were formatted in sentences where each sentence is a tuple of word and tag.

We evaluated the 2 types of models built using the BAQ Corpus as training data. The first used 3POS tags and word as features for prediction. The second type used 10POS tags and words as features. Accuracy was computed for each experiment. The average accuracy of the 10-fold cross validation was reported for each POS tagging model. Tables 2 and Table 3 shows the results of training and testing the 8NLTK's POS tagging algorithms using the BAQ Corpus.

The second experiment evaluates the previous POS tagging models on a selected sample of MSA text. This experiment was designed and implemented to test our hypothesis that Qur'an can be a good gold standard for Arabic NLP applications. The results of testing each POS tagging algorithm are shown in Table 3.

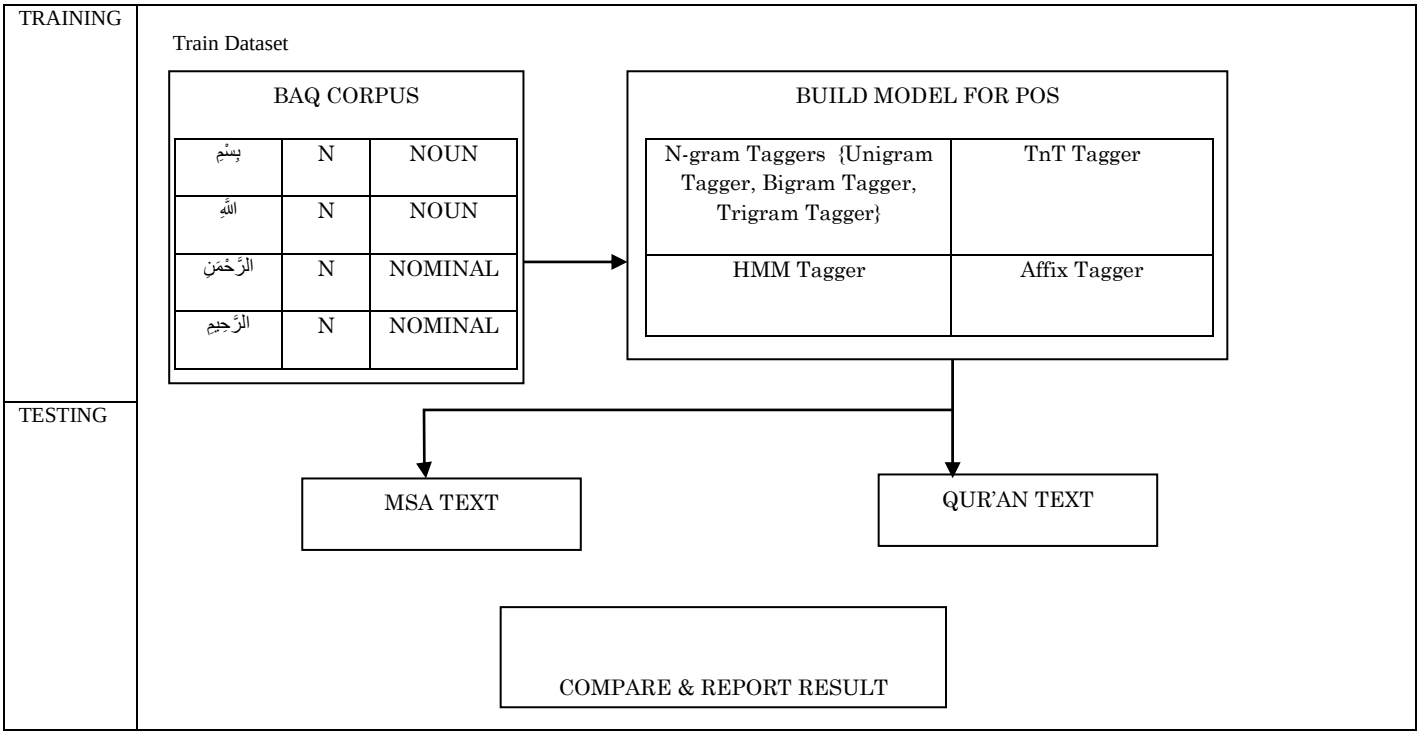

Fig. 1: Methodology for Building Stochastic POS Tagging Models for Arabic Text

Table 2: Accuracy for POS taggers using 3POS and words as features

S.NO	Tagger	Accuracy for 3POS
1	Baseline accuracy	0.5264
2	Unigram Tagger	0.9069
3	Bigram Tagger	0.9079
4	Trigram Tagger	0.9078
5	N-gram Tagger"fourth"	0.9074
6	TNT Tagger	0.907
7	Prefix Tagger	0.8745
8	Suffix Tagger	0.8339
9	Brill Tagger	0.9076
10	HMM Tagger	0.8772

Table 3: Accuracy for POS taggers using 10 POS and words as features

S.NO	Tagger	Accuracy for 10POS

1	Baseline accuracy	0.3722
2	Unigram Tagger	0.8860
3	Bigram Tagger	0.8893
4	Trigram Tagger	0.8891
5	N-gram Tagger"fourth"	0.8888
6	TNT Tagger	0.8889
7	Prefix Tagger	0.8285
8	Suffix Tagger	0.7413
9	Brill Tagger	0.8889
10	HMM Tagger	0.8575

The first experiment shows the baseline accuracy for POS tagging models trained and tested using BAQ Corpus scored 52.64% when 3POS and words were selected as features. The baseline POS tagger uses “N” to tag all words in the test sample. The average accuracy for all taggers scored 89.22% with 36.58% gain of accuracy. Bigram tagger scored highest accuracy of 90.79% and the Suffix tagger scored lowest accuracy of 83.39%.

In the second experiment, the baseline accuracy scored 37.22% for POS tagging models which were trained and tested using the BAQ Corpus and 10 POS tags and words as features for prediction. The baseline POS tagger uses “NOUN” to tag all words in the test samples. The average accuracy for all POS tagging models scored 86.20% with 48.98% gain in accuracy. The Bigram tagger also scored the highest accuracy of 88.39%. The suffix tagger also scored the lowest accuracy of 74.13%.

The POS tagging models built in this research performed similarly in terms of POS tagging accuracy. This is due to the corpus size for training which was relatively small and because of words in the Qur’an dataset were fully vowelized which decreases the ambiguity. The performance of POS taggers that used 3POS as features for prediction was better than POS taggers that used 10 POS tags as features.

5.2 EVALUATION OF POS TAGGING MODELS FOR MODERN STANDARD ARABIC TEXT

POS tagging models built using the Qur’an dataset (*i.e.* BAQ Corpus) were evaluated using a sample of MSA text. This experiment was designed and implemented to test our hypothesis of the Qur’an as a good gold standard for Arabic NLP applications. The MSA text sample was selected from QALB Corpus "Qatar Arabic Language Bank", as in [16]. All words in MSA text of QALB Corpus are non vowelized. This makes the text different from the fully vowelized text of the Qur’an. Therefore, it will add more challenge for the POS tagging task. The selected text sample consists of 5162 words. Original POS tags of this sample were mapped into 3 POS (noun, verb and particle) similar to POS tags in the BAQ Corpus. The results of POS tagging MSA text using POS models trained on BAQ Corpus were similar. The Bigram POS tagger scored accuracy of 69.47%.

6. CONCLUSION

The aim from this paper is to investigate and apply the applicability of NLTK’s POS tagging algorithms for CA. The Qur'an was selected as a gold standard for modeling and evaluating POS tagging algorithms for CA. The Boundary Annotated Qur’an Corpus (BAQ) was used as our main dataset for training and testing these ML algorithms for POS tagging CA text. We evaluated these POS tagging models using samples from the Quran and MSA text. The gain of accuracy of testing against Qur’an samples was 36.56% for POS models where 3POS tags were selected as features for prediction and 48.98% gain in accuracy for models that use 10POS as features.

7. References

[1] AlGahtani, Shabib, William Black, and John McNaught."Arabic part-of-speech tagging using transformation-based learning." Proceedings of

the 2nd International Conference on Arabic Language Resources and Tools, Cairo. 2009.

[2] Al-Kabi, Mohammed N., et al. "A Prototype for a Standard Arabic Sentiment Analysis Corpus." International Arab Journal of Information Technology (IAJIT) 13 (2016).

[3] Bird, Klein, and Edward Loper.Natural Language Processing with Python. "O'Reilly Media, Inc.", 2009.

[4] Dajani, Basma Ahmad Sedki. "Teaching Arabic Language: Towards a New Beginning that Stimulates Creativity." Procedia-Social and Behavioral Sciences 192 (2015): 758-763.

[5] Dukes, Kais, and NizarHabash."Morphological Annotation of Quranic Arabic." *LREC*. 2010.

[6] Ephraim, Yariv. "Hidden markov models." Encyclopedia of Operations Research and Management Science (2013): 704-708.

[7] Hasan, Fahim Muhammad, Naushad UzZaman, and Mumit Khan. "Comparison of different POS Tagging Techniques (N-Gram, HMM and Brill's tagger) for Bangla." Advances and Innovations in Systems, Computing Sciences and Software Engineering. Springer Netherlands, 2007. 121-126.

[8] Jacob Perkins. Python Text Processing with NLTK 2.0 Cookbook. Packt Publishing Ltd.2010.

[9] Khoja, Shereen. "APT: Arabic part-of-speech tagger." Proceedings of the Student Workshop at NAACL. 2001.

[10] Li, Shen, Joao V. Graça, and Ben Taskar. "Wiki-ly supervised part-of-speech tagging." Proceedings of the 2012 Joint Conference on Empirical Methods in NaturalLanguage Processing and Computational Natural Language Learning.Association for Computational Linguistics, 2012.

[11] Mohamed, Emad, and Sandra Kübler."Arabic Part of Speech Tagging."LREC. 2010.

[12] Raja, Fahimeh, Hadi Amiri, Samira Tasharofi, Mehdi Sarmadi, Hossein Hojjat, and Farhad Oroumchian. "Evaluation of part of speech tagging on Persian text." (2007).

[13] Sawalha, Majdi, Claire Brierley, and Eric Atwell. "Predicting Phrase Breaks in Classical and Modern Standard Arabic Text." *LREC*. 2012.

[14] Shamsuddin, SalahuddinMohd, and MohdZakiAbdRahman."Standard Arabic-its Historical and Originality-(In the light of Modern Linguistic Sciences)." Asia Pacific Online Journal of Arabic Studies 1.1 (2016).

[15] Tan, Ming, Wenli Zhou, Lei Zheng, and Shaojun Wang. "A large scale distributed syntactic, semantic and lexical language model for machine translation." In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1, pp. 201-210. Association for Computational Linguistics, 2011.

[16] Zaghouani, Wajdi, Nizar Habash, Ossama Obeid, Behrang Mohit, Houda Bouamor, and Kemal Oflazer. "Annotation Guidelines and Framework for Arabic Machine Translation Post-Edited Corpus." In Qatar Foundation Annual Research Conference Proceedings, vol. 2016, no. 1, p. ICTOP2013. Qatar: HBKU Press, 2016.