Hybrid Part-Of-Speech Tagger for Non-Vocalized Arabic Text

Meryeme Hadni[1], Said Alaoui Ouatik[1], Abdelmonaime Lachkar[2] and Mohammed Meknassi[1]

[1]FSDM, Sidi Mohamed Ben Abdellah University (USMBA), Morocco
[2]E.N.S.A, Sidi Mohamed Ben Abdellah University (USMBA), Morocco

ABSTRACT

Part of speech tagging (POS tagging) has a crucial role in different fields of natural language processing (NLP) including Speech Recognition, Natural Language Parsing, Information Retrieval and Multi Words Term Extraction. This paper proposes an efficient and accurate POS Tagging technique for Arabic language using hybrid approach. Due to the ambiguity issue, Arabic Rule-Based method suffers from misclassified and unanalyzed words. To overcome these two problems, we propose a Hidden Markov Model (HMM) integrated with Arabic Rule-Based method. Our POS tagger generates a set of three POS tags: Noun, Verb, and Particle. The proposed technique uses the different contextual information of the words with a variety of the features which are helpful to predict the various POS classes. To evaluate its accuracy, the proposed method has been trained and tested with two corpora: the Holy Quran Corpus and Kalimat Corpus for undiacritized Classical Arabic language. The experiment results demonstrate the efficiency of our method for Arabic POS Tagging. In fact, the obtained accuracies rates are 97.6%, 96.8% and 94.4% for respectively our Hybrid Tagger, HMM Tagger and for the Rule-Based Tagger with Holy Quran Corpus. And for Kalimat Corpus we obtained 94.60%, 97.40% and 98% for respectively Rule-Based Tagger, HMM Tagger and our Hybrid Tagger.

KEY WORDS

Part-Of-Speech Tagger, Natural Language Applications, Natural Language Parsing, Hidden Markov Model, Multi Words Term Extraction, Speech Recognition.

1. INTRODUCTION

Part-Of-Speech (POS) tagging is known as a necessary work in many areas Natural Language Processing (NLP) systems like information extraction, parsing of text and semantic processing. The POS tagging is known as assigning grammatical tags to words and symbols making a text which include a large amount of lexical information and captures the relationship between these words and their adjacent related words in a sentence, or paragraph [1][2][3].

Arabic POS Tagging is the process of identifying lexical category of the Arabic word existing in a sentence based on its context [5]. The most used categories are noun, adverb, verb and adjective. This is done on the basis of words role, both individually as well as in the sentence. Most words occurring in undiacritized Arabic text have the ambiguity in terms of their part of speech [4]. Take for example the term "ذهب", it can be treated as a noun "gold" or a verb "go".

There are three general approaches to deal with the tagging problem: Rule-based approach, Statistical approach, and Hybrid approach. The Rule-based approach consists of developing a rules knowledge base established by linguists in order to define precisely how and where to

DOI : 10.5121/ijnlc.2013.2601

assign the various POS tags. The statistical approach consists of building a trainable model and using the previously-tagged corpus to estimate its parameters. Once this is done, the model can be used to determine the tagger of other texts. Generally, successful statistical taggers are mainly based on Hidden Markov Models (HMMs). Finally, the hybrid approach consists in combining rule-based approach with a statistical one. Recently, the most of the POS Taggers use the latter approach as it gives better results.

Among the most recent works, we have favored the rule-based method proposed by A.Taani [19] over other methods for a number of reasons. First, it is simple to understand, accurate, and relies on a correct Arabic sentence structure using the metrics of syntactic patterns. Second, unlike the other taggers which are generally developed for Modern Standard Arabic (MSA) and thus may not be appropriate for the Classical Arabic (CA), the Taani's method can deal directly with the CA which as it is the language of the Holy Quran.

However, this rule-based method [19] presents some weaknesses: it may misclassify and unanalyzed some words. For example, the term "هدى" (i.e "upon, right") is not analyzed and the word "موتكم" (i.e "your death") is assigned by the incorrect "verb" tag.

To overcome these problems, we propose an Arabic Part-Of-Speech Tagging method based on hybrid approach which combines the rule-based approach with a statistical approach. The rest of this paper is organized as follows: In section 2, we describe the related works of POS tagging techniques in Arabic language. The Rule-based tagger is presented in section 3. Section 4 describes the principles of HMM tagger. Our proposed method is described in section 5. Section 6 presents experimental results. Finally, Section 7 concludes the paper and describes the future works.

2. RELATED WORKS

Part of Speech tagging is the task of labeling each word in a sentence with its appropriate syntactic category. As we have mentioned previously, there are many methods of POS tagging which can be classified in three categories: Statistical Approach, Rule-Based Approach and Hybrid Approach.

2.1. Statistical Approach: The statistical approach consists of building a trainable model and to use previously-tagged corpus to estimate its parameters, successful model during the last years Hidden Markov Models and related techniques have focused on building probabilistic models of tag transition sequences in sentence. This task is difficult for Arabic languages due to the lack of annotated large corpus. So far, numerous POs tagging methods have been presented in Arabic languages which are often statistical. Banko et al. [16] present a HMM tagger that exploits context on both sides of a word to be tagged. It is evaluated in both the unsupervised and supervised cases. Orumchian's tagger [10] is presented for Persian POS tagging which is follows the TNT POS tagger. The TNT tagger is based on Hidden Markov Models theory. This system uses 2.5 million tagged words as training data and the size of the tag-set is 38. Al Shamsi et al. [28] presented a method focuses on employing the Arabic phrase structure to overcome the problems of misclassified and not analyzed terms in Arabic text. By the phrase structure it means the valid sequence of POS tags that forms the grammatical structure of the noun and the verb phrases. Their method based on statistical approach that uses HMM to do POS tagging of Arabic text. It starts with a systematically analyzed of the Arabic language and uses a good tag set of 55 tags. Then it uses Buckwalter's stemmer to stem Arabic corpus and manually corrected any tagging errors. Finally, it builds an HMM-based model of Arabic POS tags, which will be trained on the annotated corpus. Alhadj et al. [30] propose a new method of part-of-speech tagger that can be used for analyzing and annotating traditional Arabic texts, especially the Holy Quran text.

This approach combines the morphological analysis with Hidden Markov Models (HMMs) based-on the Arabic sentence structure. The morphological analysis is used to reduce the size of the tags lexicon by segmenting Arabic words in their prefixes, stems and suffixes; this is due to the fact that Arabic is a derivational language. This analysis is conducted to determine the Arabic sentence structure by identifying the different main forms of both nominal and verbal sentences. On another hand, HMM is used to represent the Arabic sentence structure in order to take into account the linguistic combinations. Each state of the HMM is represented by a possible tag in the lexicon and the transitions between states (tags) are governed by the syntax of the sentence. Albared et al. [29] developed a Bigram Hidden Markov Model (HMM) to tackle the POS tagging problem of Arabic language. The HMM parameters are estimated from a small training data. They have studied the behaviour of the HMM applied to Arabic POS tagging using small amount of data. By using different smoothing algorithms with HMM model to overcome the data sparseness problem. The Viterbi algorithm is used to assign the most probable tag to each word in the text. Furthermore, several lexical models have been defined and implemented to handle unknown word POS guessing based on word substring i.e. prefix probability, suffix probability or the linear interpolation of both of them.

2.2. Rule-based Approach: This approach consists of developing a knowledge base of rules written by linguists to define precisely how and where to assign the various POS tags. One of them is the affix. Some affixes are proper to verbs; some are proper to nouns; and some others are used with verbs and nouns. Another, important sign in Arabic language is the pattern, which is an important guide in recognizing the word category. The approach is also used for some specific tasks. Diab et al. [15] designed an automatic tagging system to tokenize part-of-speech tag in Arabic text. Habash et al. [24] proposed a morphological analyzer for tokenizing and morphologically tagging Arabic words. Freeman [13] described an Arabic part-of-speech tagging system based on the Brill tagging system which is a machine learning system that can be trained with a previously-tagged corpus. Author used a tags set containing 146 tags extracted from Brown corpus for English language. Lee et al. [18] used a corpus of manually segmented words which appears to be a subset of the first release of the ATB (110,000 words). They obtained a list of prefixes and suffixes from this corpus which is apparently augmented by a manually derived list of other affixes. Maamouri et al. [14] presented a part-of-speech tagging system for Arabic. The authors based their work on the output of Tim Buckwalter's morphological analyzer. This tagging system is tested on a corpus consisted of 734 files extracted from the "Agence France Press" [25].

2.3. Hybrid Approach: Consists in combining Rule-Based method with a Statistical method used to assign the best tag for each of the words of input text. For hybrid methods, different Arabic taggers have recently emerged. Among these studies, Khoja [12] combines statistical and rule-based techniques and uses a tag set of 131 basically derived from the BNC English tag set. Tlili-Guiassa [17] used a hybrid method of based-rules and a memory-based learning method. One of the most recent works for POS tagging is done by Jabbari and Allison [7]. Their approach is transformation based and previously been used in English by Brill and Hepple [8, 9]. The construction of this tagger contains a trained learner machine which includes approximated rules. In fact, they applied an implementation of Error-Driven Transformation Based Learning.

Note that the most of these taggers [14] apply a translation of the Arabic original text to English language and use tags set derived from English which is not appropriate for Arabic. Other taggers [12] [13] [16] rely on a transliteration of the Arabic input text. Moreover, the most taggers are generally developed for Modern Standard Arabic (MSA) and thus may not be appropriate for the Classical Arabic which is the language of the Holy Quran. Generally, The hybrid methods [7] [8] [9] gives the better results for POS tagging.

Among the most recent works, we have chosen the rule-based method presented by A.Taani [19] because it is accurate and relies on correct Arabic sentence structure using the metrics of syntactic patterns. Moreover, it uses the Classical Arabic (CA) which is the language of the Holy Quran. However, this rule-based method [19] presents some problems: it may misclassify and not analyze some words. For example, the term "هدى" (i.e "upon" or "right") is not analyzed and the word "موتكم" (i.e "your death") is assigned by the incorrect "verb" tag. To resolve these problems, this paper proposes an Arabic Part-Of-Speech Tagging method which combines the Rule-Based approach [19] with HMM technique. The latter is superior to other models in term of training time and is suitable for application dealing with large amounts of text. In the following two sections, both Taani's Rule-based Tagging and HMM-based Tagging methods will be presented and detailed.

3. TAANI'S RULE-BASED TAGGING METHOD

The Taani's Rule-Based tagging method [19] allows labelling the words in a non-vocalized Arabic text to their tags. It is constituted of three main phases: the lexicon analyzer, the morphological analyzer, and the syntax analyzer. Figure 1 shows the architecture of this system.
Lexicon Analyzer: In this step, a lexicon of stop lists in Arabic language is defined. This lexicon includes prepositions, adverbs, conjunctions, interrogative particles, exceptions and interjections. All the words have to pass this phase. If the word is found in the lexicon, it is considered as tagged. Else, it passes to the next step.

Morphological Analyzer: Each word which has not been tagged in the previous phase will immigrate to this phase. A set of the affixes of each word are extracted. An affix may be a prefix, suffix or infix. After that, these affixes and the relations between them are used in a set of rules to tag the word into its class. Note that this phase is the core of the system, since it distinguishes the major percentage of untagged words into nouns or verbs.

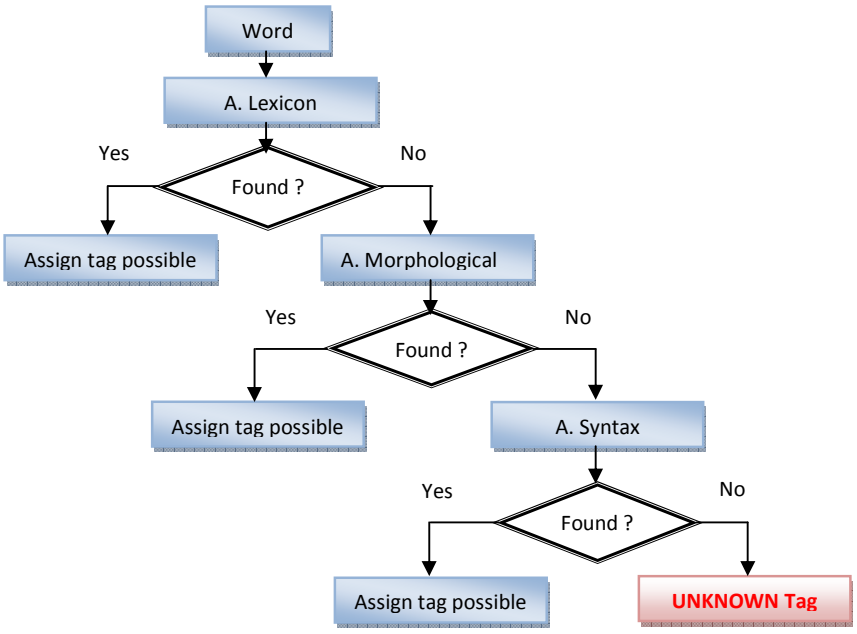

Figure1. Architecture of the Rule-Based Arabic POS Tagger [19]

Syntax Analyzer: This phase can help in tagging the words which the previous two phases failed to tag. It consists of two rules: sentence context and reverse parsing. The sentence context rule is based on the relation between the untagged words and their adjacent. Arabic language has some types of relations between adjacent words. For example the preposition and interjections are always followed by nouns. These relations may allow tagging the words into its corresponding classes. The reverse parsing rule is based on Arabic context-free grammar. The authors propose a set of rules which are used frequently in Arabic language.

In the following section, we present the HMM model since it will be integrated in our method for POS tagging Arabic text.

4. HIDDEN MARKOV MODEL

This section covers the use of a Hidden Markov Model (HMM) to do part-of-speech tagging can be seen as a special case of Bayesian inference [20]. It can be formalized as follows: for a given sequence of words, what is the best sequence of tags which corresponds to this sequence of words? If we represent an entered text (sequence of morphological units in our case) by $W = (w_i)_{1 \le i \le n}$ and a sequence of tags from the lexicon by $T = (t_i)_{1 \le i \le n}$, we have to compute:

$$\max_T [P(T|W)] \quad (1)$$

By using the Bayesian rule and then eliminating the constant part, $P(T)$ the equation can be transformed to this new one:

$$\max_T [P(T|W) * P(T)] \quad (2)$$

Where $P(T)$ represents the probability of the tag sequence (tag transition probabilities), and can be computed using an N-gram model, as follows:

$$P(T = t_1 t_2 \dots t_n) = \prod_{i=1}^{n} P(t_i | t_{i-n} \dots t_{i-2} t_{i-1}) \quad (3)$$

A tagged training corpus is used to compute $P(t_i | t_{i-n} \dots t_{i-2} t_{i-1})$, by calculating frequencies of N-gram as follows:

$$P(t_i | t_{i-n} \dots t_{i-2} t_{i-1}) = (f(t_{i-n} \dots t_{i-2} t_{i-1}) | f(t_{i-n} \dots t_{i-2} t_{i-1})) \quad (4)$$

However, it can happen that some trigrams (or bigrams) will never appear in the training set; so, to avoid assigning null probabilities to unseen trigrams (bigrams), we used a deleted interpolation developed by [20]:

$$\lambda_1 P(t_i | t_{i-n} \dots t_{i-n-1}) + \dots + \lambda_{n-2} P(t_i | t_{i-2} t_{i-1}) + \lambda_{n-1} P(t_i | t_{i-1}) + \lambda_n P(t_i) \quad (5)$$

Where $\lambda_1 + \lambda_2 + \dots + \lambda_n = 1$

Then, for calculating the likelihood of the word sequence given tag $P(T|W)$, the probability of a word appearing is generally supposed to be dependant only on its own part-of-speech tag. So, it can be written as follows:

$$P(W|T) = \prod_{i=1}^{n} P(w_i|t_i) \quad (6)$$

In addition, a tagged training set has to be used for computing these probabilities, as follows:

$$P(w_i|t_i) = \frac{f(w_i,t_i)}{f(t_i)} \quad (7)$$

Where $f(w_i,t_i)$ and $f(t_i)$ represent respectively how many times w_i is tagged as t_i and the frequency of the tag t_i itself.

Tag sequence probabilities and word likelihoods represent the HMM model parameters: transition probabilities and emission (observation) probabilities. Once these parameters are set, the HMM model can be used to find the best sequence of given a sequence of input words. The Viterbi algorithm can be used to perform this task.

We have favored to compare our tagger with Albared's HMM [29] over other methods because it is recent and these result are very encouraging especially with a small size of training corpus.

5. OUR METHOD FOR ARABIC POS-TAGGING

The proposed POS Tagger for Arabic Text is based on hybrid approach; it combines the Rule-Based method presented by Taani's [19] with a HMM model (see Figure 2). As we have mentioned, the Rule-based method is composed of three steps: lexicon analyzer, morphological analyzer and syntax analyzer.

Almost all words are recognized by rule-based method. However, some terms are not analyzed or misclassified. These terms (the rest failed terms) will be analyzed using the HMM model. The states of this model correspond to part-of-speech tags and the observations correspond to words (see Figure 3). The basic idea of our HMM model which adopts the supervised learning is to assign the most probable tag to the word of an input sentence. Two major steps are required: the training step and the test step.

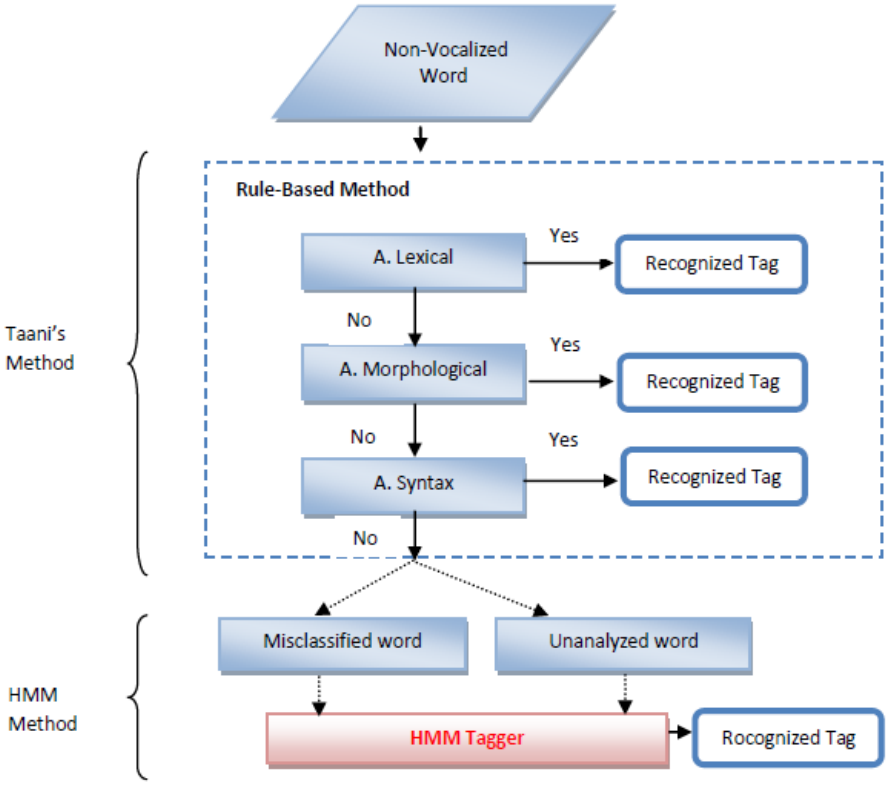

Figure2. Flowchart of the Proposed Arabic POS Tagger

The Training Step is based on supervised learning. It allows to learn the parameter of the HMM model using the corpus by estimating the transition and emission probabilities. First, for each iteration (concerning one term) we compute the emission probability for each tag i.e. p (word $|tag_i$) (see equation 7). Second, for each iteration (concerning one tag) we calculate the transition probabilities which represent the relation between tag and previous tag i.e. p (tag| previous tag) (see equation 7) for the Hidden Markov Model. The results of this step are two matrices: the matrix of transition probabilities (Tag/ Tag) and the matrix of emission probabilities (word/Tag).

The Testing Step aims to assign the best probable word's tag for which the term has been misclassified or unanalyzed during the rule-based process. First, we give all possible tags for this word. Then, we compute the probabilities of each tag of this word by using the transition probabilities and emission probabilities. Finally the Viterbi algorithm is used to calculate the best probable path (best tag sequence) for a given word in a sequence (sentence).

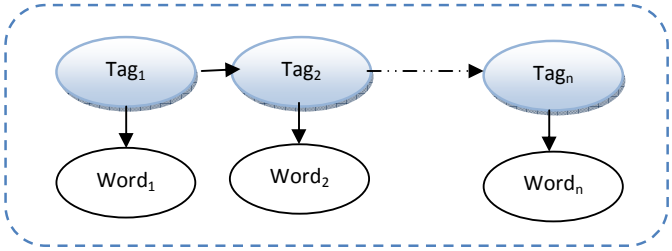

Figure 3: The Architecture of HMM Tagger

To illustrate these steps, we are considering as example the following sentence .

ذلك الكتاب لا ريب فيه هدى للمتقين
This is the Book about which there is no doubt, guidance for those conscious of Allah

Figure 4: Example of sentence including an Arabic word with unknown tag

For the word "هدى" its tag is unknown by the rule based method. First, we assign it the three different tags N, V, and INL. Using the HMM model, the transition and emission probabilities previously estimated in the training phase, we compute the probabilities of each tag for the word"هدى" as the function of probabilities of previous tag in the sentence:

- P (هدى\N) = P(هدى\N) P(N\P) P(P\N) P(N\P) P(P\N) P(N\P)
- P (هدى\V) = (هدى\V) P(N\P) P(P\N) P(N\P) P(P\N) P(N\P)
- P (هدى\INL) = (هدى\INL) P(N\P) P(P\N) P(N\P) P(P\N) P(N\P)

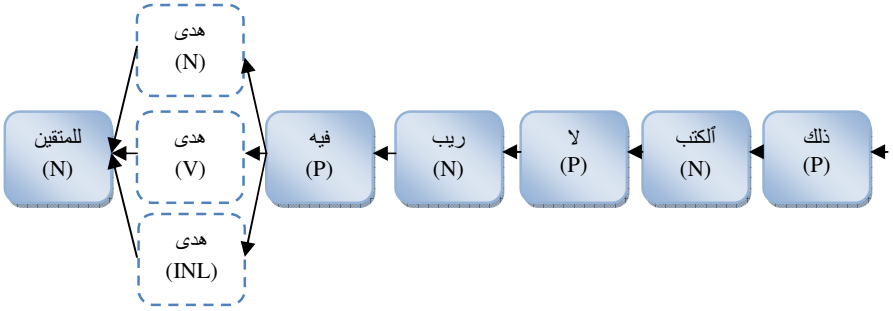

Figure 5: Example of sentence including an Arabic word with unknown tag

Finally, we calculate the best probable path (i.e. best tag sequence) by using the Viterbi algorithm. The latter is the most common decoding algorithm for HMM that gives the most likely tag sequence given a set of tags.

6. RESULTS AND DISCUSSIONS

In this section, we present and describe the used corpora to evaluate our proposed method for Arabic POS tagging. Then, we present some pre-processing task done on the corpora, and describe the tag set that we used. Experimental results will be presented and discussed.

6.1. Corpora Description

The experiments are conducted using two corpora: the Quranic Arabic Corpus [22] named as:"quranic-corpus-text-0.2" and the Kalimat Corpus [26]. The language form of these corpora is classical Arabic (CA). The Arabic alphabet of the corpus consists of the following letters: أ ب ت ث

ا ب ت ث ج ح خ د ذ ر ز س ش ص ض ط ظ ع غ ف ق ك ل م ن ه و ي ء ى ة ؤ ئ لا. The phonetic transcription for letters is shown in Table1.

6.1.1. Holy Quran Corpus

The Holy Quran corpus consists of 6236 sentences with total of 77430 words and used 33 tags [23]. This tags set consists of the following : Noun (N), Proper Noun (PN), Number (NUM), Adjective (ADJ), Imperative verbal noun (IMPN), Verb (V), Prohibition Particle (PRO), Negative Particle (NEG), Accusative Particle (ACC), Conditional Particle (COND), Restriction Particle (RES), Particle of Certainty (CERT), Interrogative Particle (INTG), Inceptive Particle (INC), Vocative Particle (VOC), Retraction Particle (RET), Amendment Particle (AMD), Future Particle (FUT), Exhortation Particle (EXH), Exceptive Particle (EXP), Explanation Particle (EXL), Surprise Particle (SUR), Aversion Particle (AVR), Answer Particle (ANS), Coordinating Conjunction (CONJ), Subordinating Conjunction (SUB), Time Adverb (T), Location Adverb (LOC), Personal Pronoun (PRON), Relative Pronoun (REL), Demonstrative Pronoun (DEM), Quranic Initial (INL).

Table1. Phonetic Transcription for Arabic Letters used in Holy Quran Corpus

Letter	Arabic	Transcription	Letter	Arabic	Transcription
Alif	ا	A	za	ظ	Z
Ba	ب	B	ayn	ع	‘
Ta	ت	T	Ghayn	غ	Gh
Tha	ث	Th	Fa	ف	F
Jim	ج	J	Qaf	ق	Q
Ha	ح	h	Kaf	ك	K
Kha	خ	Kh	Lam	ل	L
Dal	د	D	Mim	م	M
Dhal	د	Dh	Nun	ن	N
Ra	ر	R	Ha	ه	H
Zay	ز	Z	Waw	و	W
Sin	س	S	Ya	ي	Y
Shin	ش	Sh	Hamza	ء	‘
Sad	ص	s	Alif maksura	ى	A
Dad	ض	d	Ta marbuta	ة	T
Ta	ط	t			

The undiacritized form of the corpus is stored in a new file by removing the diacritics for all the words of the corpus. It is important to note that the undiacritized form of the Holy Quran corpus is used in few studies of NLP. Experimental results on undiacritized Arabic are useful because Arabic script is mostly written without diacritics.

Table2. Mapping from 33Tags set to 4Tags set for the Holy Quran corpus

Comprehensive tags set (33 tags)	Simplified tags set (4 tags)
N, PN, NUM, ADJ, IMPN	N
V	V
PRO, NEG, ACC, COND, RES, CERT, INTG, INC, VOC, RET, AMD, FUT, EXH, EXP, EXL, SUR, AVR, ANS, CONJ, SUB, T, LOC, PRON, REL, DEM	P
INL	INL

Another task is also done into the NLTK tools [21] in order to process the Holy Quran corpus files using simplified tag set. The simplified tag set includes only 4tags which are: Noun (N), Verb (V) Particle (P) and Quranic Initial (INL). Table 2 presents the mapping criteria used to convert the comprehensive tag set (33 tags) of the original corpus to the simplified one (4 tags).

6.1.2. Kalimat Corpus

The Kalimat Corpus is an Arabic natural language resource that consists of 20,291 Arabic articles collected from the Omani newspaper Alwatan [27] and used 33 tags. This tags set consists of the following: CC (coordinating conjunction), CD (cardinal number), DT (determiner), EX (existential there), FW (foreign word), IN (preposition/subordinating conjunction), JJ (adjective), JJR (adjective, comparative), JJS (adjective, superlative), LS (list marker), MD (modal), NN (noun, singular or mass), NNS (noun plural), NNP(proper noun, singular), NNPS (proper noun, plural), PDT (predeterminer), POS (possessive ending),PRP (personal pronoun), PRP$ (possessive pronoun), RB (adverb), RBR (adverb, comparative), RBS (adverb, superlative), RP (particle), TO (to), UH (interjection), VB (verb, base form), VBD (verb, past tense), VBG (verb, gerund/present participle),VBN (verb, past participle), VBP (verb, sing. present, non-3d), VBZ (verb, 3rd person sing. present), WDT (wh-determiner), WP (wh-pronoun).

Table3. Mapping from 33Tags set to 3Tags set for the Kalimat Corpus

Comprehensive tags set (33 tags)	Simplified tags set (3 tags)
JJ, JJR, JJS, NN, NNS, NNP, NNPS, PDT, POS,	N
VBZ, VBN, VBG, VBP, RB, RBR, RBS, VB, VBD,	V
DT, IN, PDT, CC, CD,EX, FW,LS,MD, PRP, PRP$, RP, TO, UH, WDT, WP	P

In order to process the Kalimat Corpus files using simplified tag set in the NLTK tool, the simplified tag set contains only 3tags which are: Noun (N), Verb (V) and Particle (P). Table3 presents the mapping criteria which are used to transform the comprehensive tag set (33tag) of the original corpus to the simplified one (3tags).
We implemented new Python modules to integrate the new created files of the Holy Quran Corpus and the Kalimat Corpus into the NLTK tool in order to perform our experiments.

6.2. Evaluation

For Holy Quran Corpus, We conducted several experiments using the previous corpus. The experiments are based on classical Arabic for undiacritized form according to 4 tags set. The table 4 presents some results of Holy Quran Arabic POS tagger using three Taggers: Taani's Rule-based method, Albared's HMM method and our proposed method.

Table4. Example of some obtained results using the Taani's Rule-Based tagger, Albared's HMM and our tagger for Holy Quran Corpus

Sentence	*Term*	*Taani's method*	*Albared's method*	*Proposed method*
إن/P مثل/N عيسى/ **?** عند/P ٱلله/N كمثل/N ءادم/**?** خلقه/V من/P تراب/N ثم/P قال/V لهۥ/P كن/ **?** فيكون/V	عيسى	?	N	N
	ءادم	?	N	N
	كن	N	V	V
ذلك/P ٱلكتب/N لا/P ريب/N فيه/P هدى/? للمتقين/N	هدى	?	N	N
ومن/P ٱلناس/N من/P يقول/V ءامنا/? بٱلله/N وبٱليوم/N ٱلءاخر/N وما/P هم/P بمؤمنين/?	ءامنا	?	V	V
	بمؤمنين	V	N	N
قالوا/V أتجعل/? فيها/P من/P يفسد/V فيها/P ويسفك/V ٱلدماء/N ونحن/P نسبح/? بحمدك/N ونقدس/V لك/P	أتجعل	?	V	V
	نسبح	N	V	V
وإذ/P نجينكم/V من/P ءال/? فرعون/N يسومونكم/V سوء/N ٱلعذاب/N يذبحون/? أبناءكم/N ويستحيون/? نساءكم/N وفى/P ذلكم/P بلاء/N من/P ربكم/N عظيم/N	ءال	?	N	N
	يذبحون	N	V	V
	يستحيون	N	N	V
ثم/P بعثنكم/V من/P بعد/N موتكم/? لعلكم/P تشكرون/?	موتكم	V	V	N
	تشكرون	N	N	V

Unanalyzed Term Misclassify Term Recognized Term

In this table, for each sentence, we consider only the ambiguous terms by Taani's method (misclassified and unanalyzed words). For example, the sentence constituted by 15 words:

إن مثل عيسى عند الله كمثل آدم خلقه من تراب ثم قال له كن فيكون
Indeed, the example of Jesus to Allah is like that of Adam. He created Him from dust; then He said to him, "Be," and he was.

contains one word "كن" which is misclassified and two terms (christ?) "عيسى"and "ءادم"(Adam?) which are not resolved by Taani's method. However, these terms are correctly treated by Albared's HMM and our method. In other sentence, the word "موتكم "is misclassified by Taani's method and Albared's HMM method by assigning the verb tag.

Table5. Example of some obtained results using the Taani's Rule-Based tagger, Albared's HMM and our tagger for Kalimat Corpus

Sentence	*Term*	*Taani's method*	*Albared's method*	*Proposed method*
الشفافية/N في/P المعاملات/N/المالية/N هي/P ركن/? أساس/V من/P أركان/N أي/P سوق/N مالية/N ناجحة/N	ركن	?	N	N
	أساس	V	N	N
	أي	P	N	N
	سوق	?	N	N
في/P ختام/N رائع /? و مثير/N لمهرجان/N سباق /N الهجين/N بولاية/N ثمريت/V فوز ? / المهار/N و الخريشه/N	رائع	?	V	N
	ثمريت	V	V	N
	فوز	?	N	N
عاد/V نجم/ ? منتخبنا/N الوطني/N لكرة/N القدم/N هاشم/ ? صالح/ ? إلى/P مقر/N إقامة/N اللاعبين/N بفندق/N كورديان/N	نجم	?	V	N
	هاشم	?	V	N
	صالح	?	V	N
إن/P الماضي/N ركن /? يعود/V بنا/N إلى/P بهاء/N الذكريات/N	ركن	?	N	N
	إلى	P	N	N

Unanalyzed Term Misclassify Term Recognized Term

Table 5 presents some obtained results of misclassified and unanalyzed words, achieved using Kalimat corpus.

For the sentence:

في ختام رائع و مثير لمهرجان سباق الهجين بولاية ثمريت فوز المهار و الخريشة
At the conclusion of wonderful and exciting for camel racing festival in the mandate Thamrit, the wining skill and kharichat.

The word "رائع","ثمريت" and "فوز" consider only the ambiguous terms : the word "ثمريت "which is misclassified by Albared's HMM method and Taani's method , and the term " رائع " and " فوز " which are not resolved by Taani's method, on the other side is correctly treated by our method.

Table6: Obtained accuracies of Taani's Rule-Based tagger, Albared's HMM and our POS tagger for Holy Quran Corpus and Kalimat Corpus

	Holy Quran Corpus			Kalimat Corpus		
method / **% training**	**Taani's method**	**Albared's HMM**	**Our method**	**Taani's method**	**Albared's HMM**	**Our method**
30%	94%	96%	97%	94,20%	96,40%	97,40%
70%	94,40%	96,50%	97,40%	94,40%	97,20%	97,80%
80%	94,20%	96,50%	97,60%	94,60%	97,40%	97,80%
90%	94,40%	96,80%	97,60%	94,60%	97,40%	98%

The table 6 presents the obtained accuracies using Taani's Rule-Based tagger, Albared's HMM tagger and our tagger, with different percentage values in the training step using the two Corpora: the Holy Quran and Kalimat Corpus. Our experiments illustrate that for undiacritized Arabic our tagger performed slightly better than Taani's and HMM taggers for both corpus, Kalimat and Quran. The size of the training data is increased with different variation of training corpus. The obtained results for our method achieve better performance: 94.4%, 96.80% and 97.60% for respectively Taani's method, Albared's HMM method and our method for Holy Quran Corpus. And produces an accuracy of 98% using 90% the portion of training corpus about our method, which is a very good result indeed. Generally, from 70% of the training corpus the accuracy exceeds 97.4%.

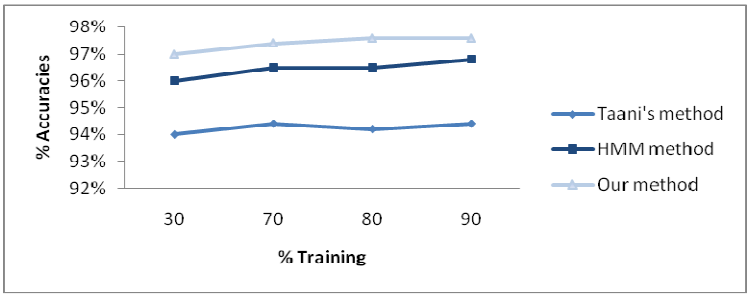

Figure5: Tagging Accuracy rate for different sizes of training corpus Kalimat of Three tagger. X-axis represents the size of the training corpus. Y-axis represents the tagging accuracy rate.

Figure 5 shows the Tagger accuracy rate as a function of training corpus size. This curve was generated by training on successive portions of our training corpus. The curve indicates that performance benefit can be obtained by increasing training set size and shows that our approach produces the good results compared to Taani's method and Albared's HMM method. Figure 5 illustrates the performance when using different sizes of Kalimat Corpus and it concludes that with the increase in the size of the corpus and using our method, the performance of the tagger also increase.

When we train the tagger on large amount of data we get accurate tagging results, we conclude that results are dependent on fraction of training data used to train the Tagger. Therefore considering the sizes of corpus used for the experiments, our tagger achieved remarkable accuracy up to 98% compared to Taani's method that has 94,60% and Albared's HMM method with a value of 97,40% for different portion of training Kalimat corpus.

7. CONCLUSION AND FUTURE WORKS

This paper describes a Hybrid POS Tagger technique for Arabic language that uses a statistical approach. The developed tagger employed an approach that combines rule-based method with Hidden Markov Models (HMMs) based-on the Arabic sentence structure. A suitable architecture of the HMM model was specified based-on the structure of sentence that allows us to deal correctly the ambiguity related to the misclassified and unanalyzed word in Arabic Rule-Based method. Having done this, two corpus composed of traditional texts of classical Arabic (CA) was used, the Quranic Arabic Corpus and the Kalimat Corpus. Parts of it were used to train and to test the tagger.

To evaluate the accuracy of the proposed POS Tagger, a series of experiments were conducted using Holy Quran Corpus and Kalimat Corpus for undiacritized Classical Arabic language. The experiments were performed with conducted further tests on more interesting dataset to evaluate the real performance of this approach. Accuracy about 98% represents a very good result of our method compared to Taani's Rule-Based method and Albared's HMM method .We note that the accuracy slightly increased with the increasing of the number of words in the training corpus. In the future, we plan to improve the tagging accuracy of unknown words by using other training corpus, and applying our POS tagger in extraction of Multi-Word Terms.

REFERENCE

[1] Lee, S.hyun. & Kim Mi Na, (2008) "This is my paper", ABC *Transactions on ECE*, Vol. 10, No. 5, pp120-122.

[2] Gizem, Aksahya & Ayese, Ozcan (2009) *Comunications & Networks*, Network Books, ABC Publishers.

[1] http://en.wikipedia.org/wiki/Part-of-speech_tagging.

[2] L.Van Guilder, (1995) "Automated Part of Speech Tagging: A Brief Overview" Handout for LING361, Georgetown University.

[3] H. Halteren, J.Zavrel & Walter Daelemans (2001).Improving Accuracy in NLP Through Combination of Machine Learning Systems. Computational Linguistics. 27(2): 199–229.

[4] DeRose & J.Steven (1990) "Stochastic Methods for Resolution of Grammatical Category Ambiguity in Inflected and Uninflected Languages." PhD.Dissertation. Providence, RI: Brown University Department of Cognitive and Linguistic Sciences.

[5] N. kumar Kumar, Anikel Dalal &Uma Sawant (2006)"hindi part of speech tagging and chunking", NLPAI machine learning contest.

[6] M. Mohseni, H. Motalebi, B. Minaei-bidgoli & M. Shokrollahi-far (2008) "A farsi part-of-speech tagger based on markov". In the proceedings of ACM symposium on Applied computing, Brazil.

[7] S. Jabbari &B. Allison(2007)"Persian Part of Speech Tagging", In the Proceedings of Workshop on Computational Approaches to Arabic Script-Based Languages (CAASL-2), USA.

[8] E. Brill (1995) "Transformation-Based Error-Driven Learning and Natural Language Processing: A case Study in Part of Speech Tagging", Computational Linguistics, USA.

[9] M. Hepple (2000), "Independence and Commitment: Assumptions for Rapid Training and Execution of Rule-based Part of-Speech Taggers", In Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (ACL). Hong Kong.

[10] T. Brants (200),"TNT – a Statistical Part-of-Speech Tagger", In the Proceedings of 6th conference on applied natural language processing (ANLP), USA.

[11] K. Megerdoomian (2004), "Developing a Persian part-of speech tagger", In the Proceedings of first Workshop on Persian Language and computer, Iran .

[12] Khoja, S.(2001) " APT: Arabic part-of-speech tagger". Proceeding of the Student Workshop at the 2nd Meeting of the NAACL, (NAACL'01), Carnegie Mellon University, Pennsylvania, pp: 1-6. http://zeus.cs.pacificu.edu/shereen/NAACL.pdf

[13] Freeman A (2001), "Brill's POS tagger and a morphology parser for Arabic", In ACL'01 Workshop on Arabic language processing.

[14] Maamouri M, Cieri C. (2002). "Resources for Arabic Natural Language Processing at the LDC", Proceedings of the International Symposium on the Processing of Arabic,Tunisia, pp.125-146.

[15] Diab M., Hacioglu K. and Jurafsky D. (2004), "Automatic Tagging of Arabic Text: From Raw Text to Base Phrase Chunks". proc. of HLTNAACL'04: 149–152.

[16] Banko M, Moore R. C. (2004). "Part of Speech Tagging in Context", Proc of the 20th international conference on Computational Linguistics, Switzerland.

[17] Tlili-Guiassa Y. (2006) "Hybrid Method for Tagging Arabic Text". Journal of Computer Science 2 (3): 245-248.

[18] L. Young-Suk, K. Papineni & S. Roukos (2003), "Language Model Based Arabic Word Segmentation," in Proceedings of the Annual Meeting on Association for Computational Linguistics, Japan, pp. 399- 406.

[19] A.T Al-Taani & S. Abu-Al-Rub (2009),"A rule-based approaches for tagging non-vocalized Arabic words". The International Arab Journal of Information Technology, Volume6 (3): 320-328.

[20] T. Brants (2000)," TnT: A statistical part of speech tagger", Proceedings of the 6th Conference on Applied Natural Language Processing, Apr. 29- May 04, Association for Computational Linguistics Morristown, New Jersey, USA., pp: 224-231.

[21] NLTK, Natural Language Toolkit. http://www.nltk.org/Home

[22] Quranic Arabic Corpus: http://corpus.quran.com

[23] Quran Tagset: http://corpus.quran.com/documentation/tagset.jsp

[24] N. Habash & O. Rambow (2005), "Arabic Tokenization, Part-of-Speech Tagging and Morphological Disambiguation in One Fell Swoop," in Proceedings of the Annual Meeting on Association for Computational Linguistics, Michigan, pp. 573-580.

[25] http://sibawayh.emi.ac.ma/web/s/?q=node/79

[26] http://bit.ly/16jO3Ks

[27] http://www.alwatan.com/

[28] F. Al Shamsi & A.Guessoum(2006)," A Hidden Markov Model–Based POS Tagger for Arabic", 8es Journées internationales d'Analyse statistique des Données Textuelles (JADT).

[29] M. Albared & O.Nazlia(2010)," Automatic Part of Speech Tagging for Arabic: An Experiment Using Bigram Hidden Markov Model ",Springer-Verlag Berlin Heidelberg, LNAI 6401, pp. 361–370.

[30] Y.O. Mohamed Elhadj(2009)," Statistical Part-of-Speech Tagger for Traditional Arabic Texts", Journal of Computer Science 5 (11): 794-800.

Authors

Miss. Meryeme Hadni Phd Student in Laboratory of computer and Modelization, Faculty of Sciences, University Sidi Mohamed Ben Abdellah (USMBA), Fez, Morocco. She has also presented different papers at different National and International conferences.

Pr. Abdelmonaime LACHKAR : received his PhD degree from the USMBA, Morocco in 2004, He is Professor and Computer Engineering Program Coordinator at (E.N.S.A, FES), and the Head of the Systems Architecture and Multimedia Team (LSIS Laboratory) at Sidi Mohamed Ben Abdellah University, Fez, Morocco. His current research interests include Arabic Natural Language Processing ANLP, Arabic Web Document Clustering and Categorization, Arabic Information Retrieval Systems, Arabic Text Summarization, Arabic Ontologies development and usage, Arabic Semantic Search Engines (SSEs).

Pr. Said Alaoui Ouatik i s working as a Professor in Department of Computer Science, Faculty of Science Dhar EL Mahraz (FSDM), Fez, Morocco. His research interests include high-dimensional indexing and content-based retrieval, Arabic Document Categorization. 2D/3D Shapes Indexing and Retrieval in large 3D Objects Database.

Mohammed Meknassi received Ph. D degree in computer sciences from Montreal University in 1993. Since 1993, he is professor of computer sciences. He teaches and makes his scientific research in the following fields: Parallel processing, Distributed Computing, Operating Systems and Image Processing. He is a member of the research unit: Systems Image and Multimedia (SIM) attached to the laboratory: Computer Sciences, Statistics and Quality (LISQ). He is the chief of the computer Sciences Department in the Faculty of Sciences Dhar El Mahraz of Fez.