Hidden Markov Model Tagger for Applications Based Arabic Text: A review

Jabar H. Yousif

***ABSTRACT*- The immense increase in the use of the Arabic Language in transmitting information on the internet makes the Arabic Language a focus of researchers and commercial developers. The developing of an efficient Arabic POS tagger is not an easy task due to the complexity of the Language itself and the challenges of tagging disambiguation and unknown words. This paper aims to explore and review the use of Part of speech Tagger for Arabic text based on Hidden Markov Model. Besides, it is discussed and explored the implementation of POS tagger for different languages. This study examined a group of research papers that applied the Part of Speech to Arabic using the Hidden Markov Model. The results have shown that a large number of researchers achieved high accuracy rates in the classification of parts of speech correctly. Handi and Alshamsi achieved a high accuracy rate of 97.6% and 97.4% respectively. Kadim obtained an average accuracy of 75.38% for a Parallel Hidden Markov Model.**

Index Terms— Natural Language Processing, Hidden Markov Model, POS tagging, Arabic Text Applications.

1. INTRODUCTION

Natural language processing (NLP) is necessary because it provides access to and sharing vast information resources on the Internet. NLP is a method of treating the natural human language automatically or semi-automatically [1]. Corpus-based Machine Learning has been identified as a critical figure in all areas of NLP. Part of Speech (POS) tagging is a method of categorizing words according to their function, like a noun, verb, adverb, adjective, etc. NLP consists of various phases including Morphological Analysis, Syntactic and Semantic Analysis, Discourse Integration, and Pragmatic Analysis. The POS tagging occurs during the Syntactic Analysis phase, and it involves assigning words to their proper part-of-speech Tag [2]. A number of approaches have been used to address the POS tagging based supervised and unsupervised methods, as shown in Figure 1. The POS tagging has been applied to Multilanguage using various methods like Rule-Based Models [3, 4], Statistical Models [5, 6, 7], Neural Network Models [8, 9, 10, 11], and hybrid models [12, 13]. The Arabic language is essential media as it is an official language to around 250 million people, in addition to the Muslims of Arabic culture around the world [14]. Hidden Markov Model (HMM) is a statistical approach for implementing the NLP application such as morphological analysis, text classification, and Tag set categorization.

Many researchers explored and implemented a HMM as a Part of speech tagging method for a number of different languages. This paper will focus only on deploying HMM in POS tagger for Arabic Text applications as shown in table 1. Kadim [15] proposed two models for Part of speech tagger based Hidden Markov Models working in parallel using the Nemlar Arabic corpus. The second model is used as a reference for determining the text with low probability tags. They also proposed a new tag-set for deploying diacritics and grammatical rules of the Nemlar Arabic corpus. The experiments used 40 sentences with 485 words, which achieved an accuracy of 98.22% for the Tagger1, an accuracy of 75.12% for Tagger2 and an accuracy of 75.38% for Tagger2 with Parallel HMM model. Alhasan [16] introduced an active part of speech tagging method for the Arabic language called Bee Colony Optimization algorithm. They tested and evaluated the proposed POS tagger using KALIMAT corpus with 18M words. The proposed tagger achieved an accuracy of 98.2. Khetam [17] investigated the performance of various types of POS tagger for Arabic text-based different corpus (BAQ, QALB). Multiple experiments were performed for testing the accuracy of POS tagging models trained using BAQ Corpus. They achieved 89.22% average of accuracy for all taggers. Besides, the Bigram tagger obtained an accuracy rate of 90.79%, and the Suffix tagger scored an accuracy of 83.39%. Ba-Alwi [19] performed a comparison study between the morpheme-based and word-based statistical POS tagger approaches. They examined the performance of three stochastic POS tagger models for Arabic text (HMM POS tagger based prefix guessing, HMM POS tagger based linear interpolation guessing, TnT tagger). They used an annotated linguistic corpus called Quranic Arabic, which present and define the Arabic syntax, grammar, and morphology for the words in the Holy Quran. They found that the Word-based HMM-POS tagger achieved an accuracy rate of 88.1%, and the Morpheme-based TnT-POS obtained an accuracy rate of 93.8%. Zeroual [19] presented a probabilistic Part of speech (POS) tagger for Arabic text based on Hidden Markov Models (HMM) called Tree Tagger. Besides, they proposed comprehensive tag-set, which can be used for 22 different languages, including both Modern Standard Arabic and Classical Arabic. The proposed tagger obtained accuracy rates of 99.4% using Al-Mus’haf corpus. It also achieved an accuracy of 93.86% using the NEMLAR corpus. Ahmed [20] utilized a

Jabar H. Yousif , Faculty of Computing and Information Technology, Sohar University, Oman; jyousif@soharuni.edu.om

combined three taggers named HMM, Brill, and Max match taggers to proposed new tagger called Master-Slave tagger. It is implemented using a data set of 29k words. The experimental results show that the combined master-slaves tagger achieved an accuracy of 90.05%. Also, they improved the performance, after adding a rule-based tagger, to a rate of 92.86 %. The tokenizing words' accuracy is 98.8%.

Hadni [21] introduced a hybrid POS tagger with 33 tag-set based Hidden Markov Model (HMM) and Rule-Based. They used the Holy Quran Corpus and achieved an accuracy rate of 97.6% and 96.8% respectively. Also, they used Kalimat Corpus and scored 94.60% and 97.40 sequentially. Aajmi [22] presented a new POS tagger method based on Hidden Markov Model (HMM) for extracting word weights by removing the prefixes and suffixes attached to a word. The proposed method achieved a promising accuracy of 95 %. Köprü [23] discovered a fast and straightforward Part of Speech tagger based HMM for Arabic text. They obtained an accuracy of 95.57% using standard tag-set for Arabic.

AbuZeina [24] proposed Part of Speech tagger based Stanford Arabic tagger for utilizing an Arabic speech recognition system. They achieved an accuracy of 90.18% using 5.4 hours speech corpus of standard Arabic. Mohamed [25] performed a comparison study between two part of speech tagging for Arabic text with and without word segmentation using the Penn Arabic Treebank tag-set. The word-based segmentation POS tagger achieved the best accuracy results of 94.74%. Albared [26] implemented a Bigram Hidden Markov Model for deploying the POS tagging for Arabic text. They process the unknown words by extracting the stem of the word and trying to remove prefix and suffix attached to the stem. The experiments have shown that the achieved accuracy is 95.8%. El Hadj [27] introduced a Part of Speech Tagging for Arabic language using a combination of HMM and morphological analyzer. The POS tagger obtained an accuracy rate of 96%. Al Shamsi [28] proposed a new approach to Part of Speech tagging for extracting the Named Entities using 10 MBs of Arabic corpus with small tag-set consists of 55 tags. They achieved an accuracy rate of 97%.

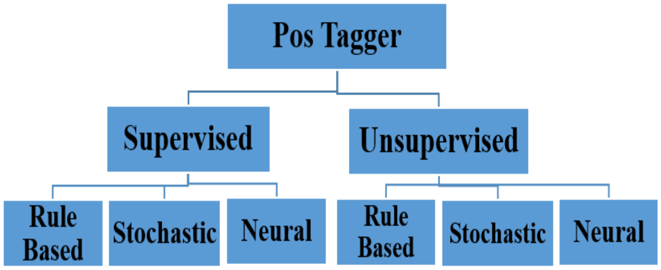


Figure 1: Implementation of Part of Speech Tagger

2. METHODS & EVALUATION FACTORS

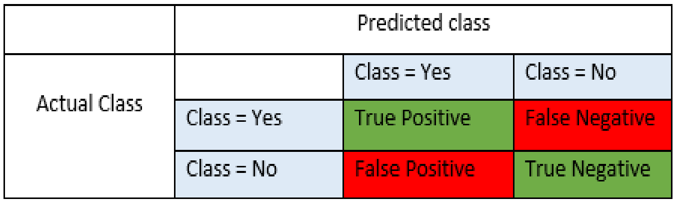
2.1 Examining the tagger errors and performance

Several factors could affect the taggers' results, which includes training data, vocabulary syntactical and grammatical styles, and the method used in the classification of tags. The tagger errors can be categorized into three main classes as follows:

- Failures due to the insufficient data for training process, which used to estimate the model outputs. The errors in predicting the output (category of word) is happened because of the unidentified output or inaccurately defined the conditions of word probabilities.
- Failures due to the error in defining the grammatical styles of the testing text. These errors are appeared because the testing data contains unknown grammar that not comprised in the training data sets. Therefore, using previous trained models to minimize the rate of errors.
- Failures due to the inadequate of model hypotheses, which describes a general question or statement that suggests a possible association between two or more objects.

Many methods are used to examine the performance of POS tagging. However, the simplified scheme of evaluating the correctness of the classification tag is the accuracy. In addition, it is essential to highlight some factors that could affect the accuracy of tagging like the type of vocabulary, Tag set size, Test corpus size, and the method used to compute the production. The accuracy is the number of correct answers divided by the total number of responses (both correct and incorrect).

Also, recall, Precision, and F-measure are used in evaluating the system performance, which is illustrated in Figure 2. True Positives (TPov) are correctly predicted the actual results with positive output. Also, the True Negatives (TNeg) are correctly predicted the actual results with positive values. In addition, the False Positives (FPov) and False Negatives (FNeg) are wrongly predicted the real values [29].



		Predicted class	
Actual Class		Class = Yes	Class = No
	Class = Yes	True Positive	False Negative
	Class = No	False Positive	True Negative

Figure 2: Factors to compute recall, Precision and F-measure

These factors are defined as in the following equations:

Accuracy = TPov+TNeg/TPov+FPov+FNeg+TNeg … (1)

Precision = TPov/TPov+FPov … (2)

Recall = TPov/TPov+FNeg … (3)

F1 Score = 2*(Recall*Precision) /(Recall +Precision)… (4)

Table 1: comparison of different research papers implementing HMM of POS tagging for Arabic Text

Ref. No./Year	Location	Method	Corpus / Tag-Set	Accuracy
[15] Kadim /2018	Morocco	HMM- Viterbi algorithm	Nemlar Arabic Corpus,	75.38
[16] Alhasan / 2018	Jordan	HMM- Bee Colony Optimization algorithm.	Kalimat Corpus 18M words.	98.2
[17] Khetam / 2017	Jordan	HMM- Bigram tagger	BAQ Corpus	90.79
[18] Ba-Alwi /2017	Yemen	HMM with prefix guessing, TnT tagger	Quran Arabic Corpus	88.1 93.8%
[19] Zeroual / 2016	Qatar	HMM- Viterbi algorithm	Al-Mus'haf Corpus, NEMLAR Corpus	99.4 93.86
[20] Ahmed /2013	Poland	HMM (Master and Slaves)	Multilevel Arabic tag-set	92.86
[21] Hadni M. /2013	Tunisia	Hybrid HMM +Rule Based	Holy Quran , Kalimat Corpus	97.6
[22] Alajmi / 2011	Egypt	HMM	15 M. W.	95
[23] Köprü / 2011	Turkey	HMM	Standard Arabic Tag-set	95.57
[24] AbuZeina /2011	KSA	HMM Stanford tagger	5.4 H. speech Corpus	90.18
[25] Mohamed /2010	USA	HMM	Penn Arabic Treebank	94.74
[26] Albared /2010	Malaysia	HMM- Viterbi algorithm	Arabic Corpus of 23146 W.	95.8
[27] El Hadj / 2009	KSA	HMM	Manual corpus of old text.	96
[28] Al Shamsi / 2006	UAE	HMM	10 MBs of Arabic Corpus	97

$$T_0 = \arg_T \max P(T \mid W) = \arg_T \max \frac{P(W \mid T) * P(T)}{P(W)} = \arg_T \max P(W \mid T) * P(T) \qquad \ldots 5$$

2.2 Hidden Markov Model (HMM)

HMM is a finite set of states machine based on a statistical Markov method used as a classifier, which involves a set of events with specific input values. The probability of each event occurring depends on the state achieved in the previous state [30].

Let Xn and Yn are two stochastic discrete time events and n > 0 , then the association (Xn,Yn) is called HMM if :

$- Xn\ is\ markon\ process\ and\ not\ hiddden.$

$- P(A(Yn \in A|X1 = x1, \cdots Xn = xn)$

$= P(A(Yn \in A|Xn = xn)\ , n \geq 1$

Where Xn is called the hidden state, and $P(A(Yn \in A|Xn = xn)$ is the output probability.

Consider a stochastic sequence of tags is (T) that should be assigned to the number of words in a sentence (W). Then we can compute the function of both lexical probability of P(W|T) and the probability of the language model P(T) using Bayes' rule as defined in equation 5. The transition matrix consists of all possible transitions from the current state (input) to next state (output). For example, consider we have three Part of Speech tags (Noun, Verb, and Adverb). Figure 3 shows a Markov model with three states, and Figure 4 presents the association probability transition matrix [31].

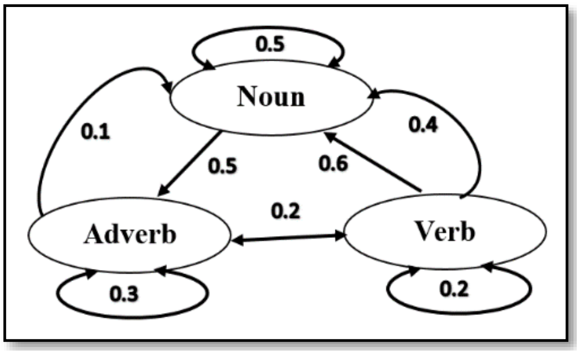
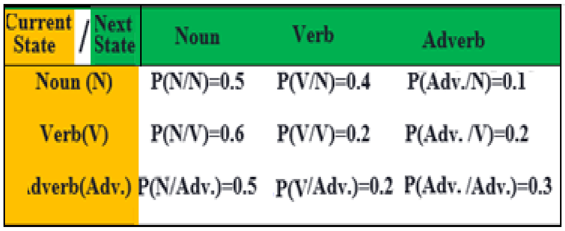


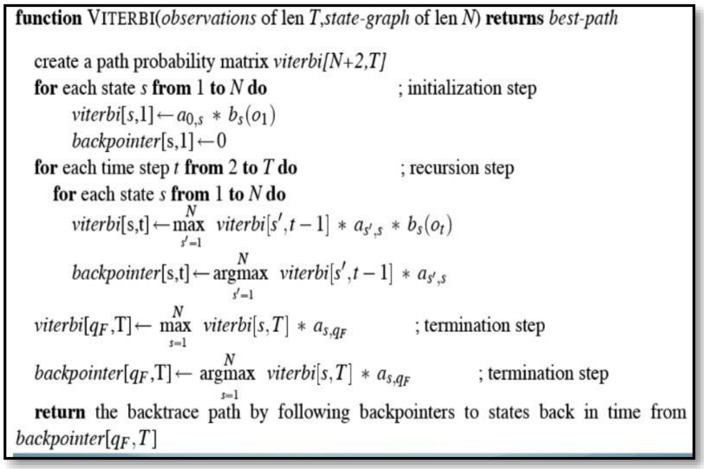
Figure 3 a Markov model with three states



Current State / Next State	Noun	Verb	Adverb
Noun (N)	P(N/N)=0.5	P(V/N)=0.4	P(Adv./N)=0.1
Verb(V)	P(N/V)=0.6	P(V/V)=0.2	P(Adv. /V)=0.2
dverb(Adv.)	P(N/Adv.)=0.5	P(V/Adv.)=0.2	P(Adv. /Adv.)=0.3

Figure 4 the association probability transition matrix.

For the sake of implementing HMM as tagger for as a sequence of words, then the probability of current word N will compute based on the probability of previous word N-1. This approach is called the first-order Markov model (Bigram). Besides, another method called second-order Markov model (Trigram), in which the probability of current word N will be computed based on the probability of the previous two words N-1and word N-2. The Stochastic hypothesis should help in determine probability of the unidentified word's P(W|T); to use the forward-backward and the Viterbi algorithm, as illustrated in Figure 5 [32]. However, the researchers proposed several methods for estimating the probabilities of word morphology.



```
function VITERBI(observations of len T, state-graph of len N) returns best-path

  create a path probability matrix viterbi[N+2,T]
  for each state s from 1 to N do                          ; initialization step
      $viterbi[s,1] \leftarrow a_{0,s} * b_s(o_1)$
      $backpointer[s,1] \leftarrow 0$
  for each time step t from 2 to T do                      ; recursion step
     for each state s from 1 to N do
      $viterbi[s,t] \leftarrow \max_{s'=1}^{N} viterbi[s',t-1] * a_{s',s} * b_s(o_t)$
      $backpointer[s,t] \leftarrow \operatorname{argmax}_{s'=1}^{N} viterbi[s',t-1] * a_{s',s}$
  $viterbi[q_F,T] \leftarrow \max_{s=1}^{N} viterbi[s,T] * a_{s,q_F}$          ; termination step
  $backpointer[q_F,T] \leftarrow \operatorname{argmax}_{s=1}^{N} viterbi[s,T] * a_{s,q_F}$      ; termination step
  return the backtrace path by following backpointers to states back in time from
backpointer[$q_F$, T]
```

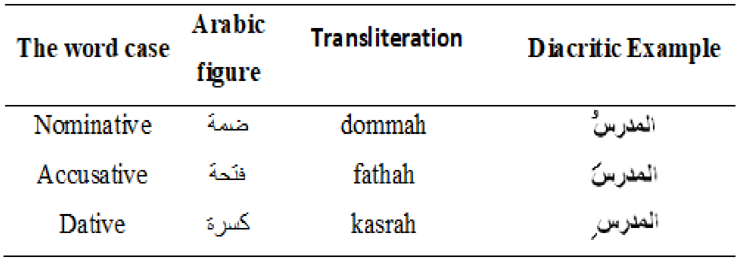
Figure 5: Viterbi Algorithm architecture Source [32]

3. ARABIC LANGUAGE SPECIFICATION & CORPUS

3.1 Arabic Language characteristics

Arabic morphology is complicated compared to English and European languages, as it comprehends different patterns of morphology called *root*. The meaning can be drawn from the consonants in the root (Anne 2000). The syntactic functions can be concluded from the vowels on the last letter with free-position, which are frequently neglected in written Arabic. Furthermore, the writing system does not reveal the boundaries of words quickly by spaces as other languages like English. The difficulty of analyzing Arabic text is notorious among Arabs even for the well-trained human, let alone machine.

The old Arabic language (Classical Arabic) is more standardized in the new figure compared to Modern Standard Arabic (MSA). Classical Arabic is used in several countries as official media of publications, writing, and speaking. The Arabic words are highly derivative and inflective in nature and structure, which are considered as phrases rather than single words. The Arabic alphabet consists of 28 letters, and the writing is directed from right to left in horizontal lines. The Arabic character has a diacritic figure written up or down of the character to differentiate it. The diacritic determines the case of the word (nominative, accusative, dative, etc.), as clarified in Figure 6.



The word case	Arabic figure	Transliteration	Diacritic Example
Nominative	ضمة	dommah	المدرسُ
Accusative	فتحة	fathah	المدرسَ
Dative	كسرة	kasrah	المدرسِ

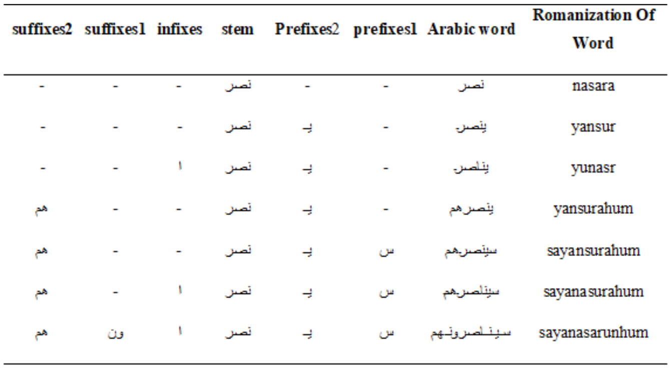
Figure 6: Arabic Diacritic with Examples

The root is the stem figure, which consists of three consonants (CCC). Words can be generated from these roots directly like in the verbs (as in the example of the past verb (درس – *drs*) that means, “studied”). Besides, affixes (infixes, prefixes, and suffixes) can be attached to the words that have been constructed from the roots to generate new words. For instance, adding the Arabic prefix "m" to the word "drs" produces the word, Arabic (مدرس – *mdrs*) which means "teacher" [1].

The Arabic word can be described as follows:

[prefixes1][prefixes1] **Stem** [infixes] [suffixes1] [suffixes2]

The prefix part is attached at beginning of the “Stem”, but the suffix part is embedded at the end. Whereas, the infixes are inserted inside the stems. Figure 7 shows some examples of the affixes handling.



suffixes2	suffixes1	infixes	stem	Prefixes2	prefixes1	Arabic word	Romanization Of Word
-	-	-	نصر	-	-	نصر	nasara
-	-	-	نصر	يـ	-	ينصر	yansur
-	-	ا	نصر	يـ	-	يناصر	yunasr
هم	-	-	نصر	يـ	-	ينصرهم	yansurahum
هم	-	-	نصر	يـ	س	سينصرهم	sayansurahum
هم	-	ا	نصر	يـ	س	سيناصرهم	sayanasurahum
هم	ون	ا	نصر	يـ	س	سيـنـاصـرونـهـم	sayanasarunhum

Figure 7 Arabic Affixes Handling

3.2 Arabic Corpus and Tag-set

A corpus (plural *corpora*) is a general term for describing a collection of data. A tagged corpus is a crucial part of automating the construction of the statistical models, which is an essential resource for both the linguistic study and Natural Language Processing. The Corpus can be defined as a large body of natural language texts. Generally, there are two main classes of Corpus called the general and annotated Corpus [33]. A tag-set is a group of classes determine the grammatical categories of parts of speech.

Many tag-sets for the Arabic language are proposed and implemented in the last decade like Khoja tag-set [34], AlQrainy tag-set [35], Sawalha tag-set [36], Alhadj tag-set [37], reduced Buckwalter tag-set [38], etc. Figure 8 shows the Khoja tag-set, which is the earliest tag-set developed by Shereen Khoja [34]. Latfia Alsulaiti [39] presented a helpful study on Arabic corpora. Many Arabic Corpora were developed and implemented [40] like AL-Hayat Corpus, Buckwalter Arabic Corpus, Nijmegen Corpus, Arabic Newswire corpus, etc.

4. RESULTS AND DISCUSSION

The research methodology used in this work included a focus on the collecting of qualitative and quantitative data through a comparative study. It is explored the methods of designing Part of Speech tagging classification algorithms based on the type and the function text in a large number of papers published in high-ranking international journals. This paper also included a comparative study of the number of research published according to the type of method used in many internationally classified databases such as Google Scholar, Science Direct, and Medley. Besides, this work covered a large number of natural languages such as English, Arabic, Chinese, Hindi, French, German, and other living languages.

The results of this research showed that the comparative study of the number of research published on Google Scholar to classify the Part of Speech Tagging using Hidden Markov Model (HMM) for several natural languages. Figure 9 shows that the Hidden Markov Model is the most method used for tagging Part of Speech compared to the other techniques like the rules-based and the neuronal network methods. In addition, this work explored the implementation HMM for different languages. Figure 10 illustrates the number of pubplished research papers on Google Scholar related to HMM method. It shows that the English language is the most used on the Internet by 19000 research papers applied the technique of Hidden Markov Model and followed by the Chinese language with 13000 articles and then French language with 12300 articles. Also, it indicates a lower rate number of articles that implemented related to POS using HMM for Indonesian and Bengali languages.

This paper also discussed the different ways to implement the Part of Speech tagging using the Hidden Markov Model technique in selecting the most suitable class for the type of word under execution. Many researchers were presented and discussed the process of choosing the proper tag for a word such as Bigram, Trigram, and Viterbi algorithms. The comparative study involved the results of several crucial international scholar databases such as Google Scholar, Science Direct, and Medley. Figure 11 shows that the Viterbi algorithm is commonly used in the implementation of the classification of Part of Speech tagging, followed by Bigram method and lastly Trigram method. Besides, it indicates that the Google scholar database has the highest number of research paper related to Hidden Markov Model technique based POS tagging.

Besides, this work examined the implementation of POS using HMM for different languages. Figure 12 illustrates the number of published research papers on Google Scholar related to POS tagging based on the HMM method. The results proved that the published research papers on POS using HMM related to the English language is scored the highest rate, followed by Chinese language and then both French and Japanese languages. However, the number of published research papers based on the Arabic language is scored the fifth position after the Spanish language.

Finally, this study reviewed a group of research papers that applied the Part of Speech to Arabic using the HMM method. It should be mentioned at this point that there are many reasons lead to the failure to categorize the Part of Speech in the Arabic language correctly, including the difference of Arabic idioms and type of style, whether traditional or modern. The Arabic language use dialects, which leads to giving different meaning for a word depending on the class of dialects and its location in the word. Besides, Arabic writing has complicated morphology and semantic, which affects the tagging process.

However, the Figure 13 shows that a large number of researchers achieved high accuracy rates in the classification of parts of speech correctly, including Handi [21] and Alshamsi [28] percent, while the research paper by khadim [15] achieved the lowest rating of accuracy. Hadni [21] introduced a hybrid POS tagger with 33 tag-set for tagging Arabic text extracted from the Holy Quran Corpus and Kalimat Corpus. They achieved a high accuracy rate of 97.6% and 97.4% respectively. Al Shamsi [28] proposed a Named Entities approach for extracting the tags of words. They used 10 MBs of Arabic corpus with small tag-set consists of 55 tags, which help to achieve a high accuracy rate of 97%. Also, Kadim [15] proposed parallel Part of speech tagger for Arabic text using the Nemlar Arabic corpus. They obtained an average accuracy of 75.38% for the Parallel HMM model.

5. CONCLUSION & RECOMMENDATIONS

This paper uses a qualitative and quantitative research methodology and data collection through a comparative study. Methods of designing part-speech markers have been discovered in a large number of papers published in high-profile international journals. This paper also included a comparative study of the number of research papers according to several internationally classified databases such as Google Scholar, Science Direct and Medley. In addition, this work covered a large number of natural languages such as English, Arabic, Chinese, Hindi, French, German and other living languages.

The results of this work indicates that the Hidden Markov Model is the most commonly used method of categorizing a POS tagging compared to other techniques such as rule-based methods and neural network methods. Moreover, it explains that English is the most widely used on Goggle scholar database that applied the Hidden Markov model followed by Chinese and then French. Also, it indicates a decrease in the number of POS related articles implemented using HMM for Indonesian and Bengali languages.

Also, the result shows that the Viterbi algorithm is commonly used to implement the classification of part of speech tags, followed by the Bigram method and then the Trigram method. Finally, this study shows that a large number of researchers have achieved high accuracy rates in correctly categorizing parts of speech using HMM based Arabic text with accuracy more that 97%.

The recommendations of this work are:

1-There is a need for unifying the style of Arabic text (combined the traditional and modern style).
2- Unify the Arabic tag-set in a standard form that takes into consideration all variation of grammatical and syntax issues.
3- Discover an automatic method for identifying the root of a word and remove all the attached prefixes and suffixes.

REFERENCES

[1]. Yousif, J. H. (2011). Information Technology Development. LAP LAMBERT Academic Publishing, Germany ISBN 9783844316704.

[2]. Yousif, J. H. (2013). Natural language processing based soft computing techniques. International Journal of Computer Applications, 77(8).

[3]. Khoja, S. (2001, June). APT: Arabic part-of-speech tagger. In Proceedings of the Student Workshop at NAACL (pp. 20-25).

[4]. Rashel, F., Luthfi, A., Dinakaramani, A. and Manurung, R., 2014, October. Building an Indonesian rule-based part-of-speech tagger. In 2014 International Conference on Asian Language Processing (IALP) (pp. 70-73). IEEE.

[5]. Yousif, J.H. and Al-Risi, M.H., 2019. PART OF SPEECH TAGGER FOR ARABIC TEXT BASED SUPPORT VECTOR MACHINES: A REVIEW. ICTACT Journal on Soft Computing, 9(2).

[6]. Yousif, J. H., & Sembok, T. M. T. (2008, August). Arabic part-of-speech tagger based Support Vectors Machines. In Information Technology, 2008. ITSim 2008. International Symposium on Information Technology (Vol. 3, pp. 1-7). IEEE.

[7]. Diab, M., Hacioglu, K., & Jurafsky, D. (2004, May). Automatic tagging of Arabic text: From raw text to base phrase chunks. In Proceedings of HLT-NAACL 2004: Short papers (pp. 149-152). Association for Computational Linguistics.

[8]. Yousif, J.H. and Sembok, T., 2006. Design and implement an automatic neural tagger based arabic language for nlp applications. Asian Journal of Information Technology, 5(7), pp.784-789.

[9]. Yousif, J.H. and Sembok, T., 2006. Recurrent Neural Approach Based Arabic Part-Of-Speech Tagging. In proceedings of International Conference on Computer and Communication Engineering (ICCCE'06) (Vol. 2, pp. 9-11).

[10]. Yousif, J.H. and Sembok, T., 1812. Arabic Part-Of-Speech Tagger Based Neural Networks. In proceedings of International Arab Conference on Information Technology ACIT2005, ISSN (Vol. 857).

[11]. Youisf ,J. H. , 2018, Neural Computing based Part of Speech Tagger for Arabic Language: A review study , International Journal of Computation and Applied Sciences , 5(1), PP 361-365.

[12]. Vashishtha, S. and Susan, S., 2019. Fuzzy Rule based Unsupervised Sentiment Analysis from Social Media Posts. Expert Systems with Applications, p.112834.

[13]. Forsati, R. and Shamsfard, M., 2014. Hybrid PoS-tagging: A cooperation of evolutionary and statistical approaches. Applied Mathematical Modelling, 38(13), pp.3193-3211.

[14]. Abdelali, A., Cowie, J. and Soliman, H.S., 2004, April. Arabic information retrieval perspectives. In Proceedings of the 11th Conference on Natural Language Processing, Journes d'Etude sur la Parole-Traitement Automatique des Langues Naturelles (JEP-TALN) (pp. 391-400).

[15]. Kadim, A. and Lazrek, A., 2018. Parallel HMM-based approach for Arabic part of speech tagging. Int. Arab J. Inf. Technol., 15(2), pp.341-351.

[16]. Alhasan, A. and Al-Taani, A.T., 2018. POS Tagging for Arabic Text Using Bee Colony Algorithm. Procedia computer science, 142, pp.158-165.

[17]. KhetamYassen, M. and Al Zaghoul, F., 2017. PART-OF-SPEECH TAGGING FOR CLASSICAL AND MSA ARABIC TEXT USING NLTK. New Trends in Information Technology (NTIT)–2017, p.106.

[18]. Ba-Alwi, F.M., Albared, M. and Al-Moslmi, T., 2017. Choosing the Optimal Segmentation Level for POS Tagging of the Quranic Arabic. British Journal of Applied Science & Technology, 19(1), p.10.

[19]. Zeroual, I. and Abdelhak, L., 2016, March. Adapting a decision tree based tagger for Arabic. In 2016 International Conference on Information Technology for Organizations Development (IT4OD) (pp. 1-6). IEEE.

[20]. Ahmed, H.A., 2013. Arabic Morphosyntactic Raw Text part of Speech Tagging System (Doctoral dissertation, PhD dissertation, University of Warsaw).

[21]. Hadni M, Ouatik SA, Lachkar A, Meknassi M. Hybrid part-of-speech tagger for non-vocalized Arabic text. Int. J. Nat. Lang. Comput. 2013 Dec;2(6):1-5.

[22]. Alajmi, A.F., Saad, E.M. and Awadalla, M.H., 2011. Hidden Markov Model based Arabic morphological analyzer. International Journal of Computer Engineering Research, 2(2), pp.28-33.

[23]. Köprü, S., 2011, February. An efficient part-of-speech tagger for Arabic. In International Conference on Intelligent Text Processing and Computational Linguistics (pp. 202-213). Springer, Berlin, Heidelberg.

[24]. AbuZeina, D., Al-Khatib, W., Elshafei, M. and Al-Muhtaseb, H., 2011. Toward enhanced Arabic speech recognition using part of speech tagging. International Journal of Speech Technology, 14(4), p.419.

[25]. Mohamed, E. and Kübler, S., 2010, June. Is Arabic part of speech tagging feasible without word segmentation?. In Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics (pp. 705-708). Association for Computational Linguistics.

[26]. Albared, M., Omar, N., Ab Aziz, M.J. and Nazri, M.Z.A., 2010. Automatic part of speech tagging for Arabic: an experiment using Bigram hidden Markov model. In International Conference on Rough Sets and Knowledge Technology (pp. 361-370). Springer, Berlin,

Heidelberg.
DOI: 10.1007/978-3-642-16248-0_52.

[27]. Elhadj, Y.O., 2009. Statistical part-of-speech tagger for traditional Arabic texts. Journal of computer science, 5(11), p.794.

[28]. Al Shamsi, F. and Guessoum, A., 2006, April. A hidden Markov model-based POS tagger for Arabic. In Proceeding of the 8th International Conference on the Statistical Analysis of Textual Data, France (pp. 31-42).

[29]. Bach, N.X., Linh, N.D. and Phuong, T.M., 2018. An empirical study on POS tagging for Vietnamese social media text. Computer Speech & Language, 50, pp.1-15.

[30]. Forsati, R. and Shamsfard, M., 2014. Hybrid PoS-tagging: A cooperation of evolutionary and statistical approaches. Applied Mathematical Modelling, 38(13), pp.3193-3211.

[31]. Suleiman, D., Awajan, A. and Al Etaiwi, W., 2017. The Use of Hidden Markov Model in Natural ARABIC Language Processing: a survey. Procedia computer science, 113, pp.240-247.

[32]. Online Source: https://cl.lingfil.uu.se/~marie/undervisning/textanalys16/levow.pdf , Accessed 16/08/2019.

[33]. Maamouri, M., Bies, A., Buckwalter, T. and Mekki, W., 2004, September. The penn arabic treebank: Building a large-scale annotated arabic corpus. In NEMLAR conference on Arabic language resources and tools (Vol. 27, pp. 466-467).

[34]. Khoja, S., Garside, R. and Knowles, G., 2003. A tagset for the morphosyntactic tagging of Arabic. Proceedings of the Corpus Linguistics. Lancaster University (UK), 13.

[35]. Al-qrainy, S. and Ayesh, A., 2006. Developing a tagset for automated POS tagging in Arabic, WSEAS Transactions on Computers Vol 5.

[36]. Sawalha M. (2011). Open-source Resources and Standards for Arabic Word Structure Analysis: Fine Grained Morphological Analysis of Arabic Text Corpora TAGGING. PhD dissertation, School of Computing, University of Leeds, UK.

[37]. Habash, N.Y., 2010. Introduction to Arabic natural language processing. Synthesis Lectures on Human Language Technologies, 3(1), pp.1-187.

[38]. Al-Sulaiti, L. and Atwell, E.S., 2006. The design of a corpus of contemporary Arabic. International Journal of Corpus Linguistics, 11(2), pp.135-171.

[39]. Alqrainy, S., 2008. A Morphological-Syntactical Analysis Approach for Arabic Textual Tagging. 2008. Leicester, UK, De Montfort University. PhD, 197.

Dr. Jabar Yousif is working as associate Prof. at Faculty of Computing and Information Technology, Sohar University, Oman. Ph.D. Information Science & Technology, M.Sc. & B.Sc. in Computer Science. A postdoctoral fellowship to design a Holistic Future Virtual Reality Laboratory (Visi Lab). I have more than 20 years of teaching experience. I published more than 70 papers & books in the fields of Artificial Intelligent, Renewable Energy, Cloud Computing, Soft Computing, Artificial Neural Networks, Natural Language Processing, Arabic Text Processing, Virtual Reality. Editorial board & reviewer for many scientific journals and conferences.

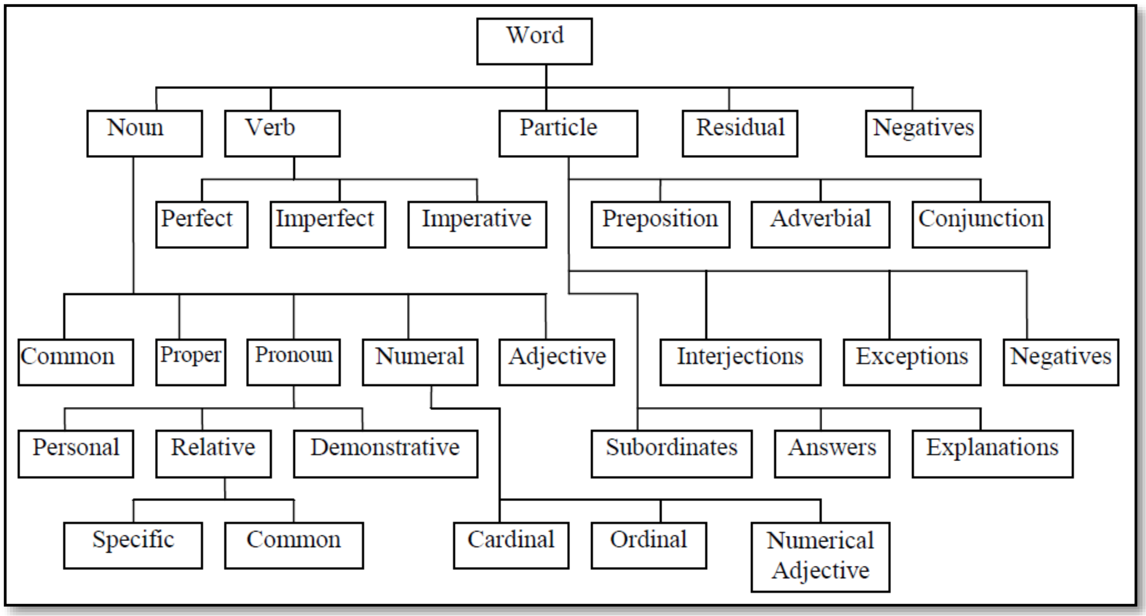


Figure 8: Khoja tag-set hierarchy [34

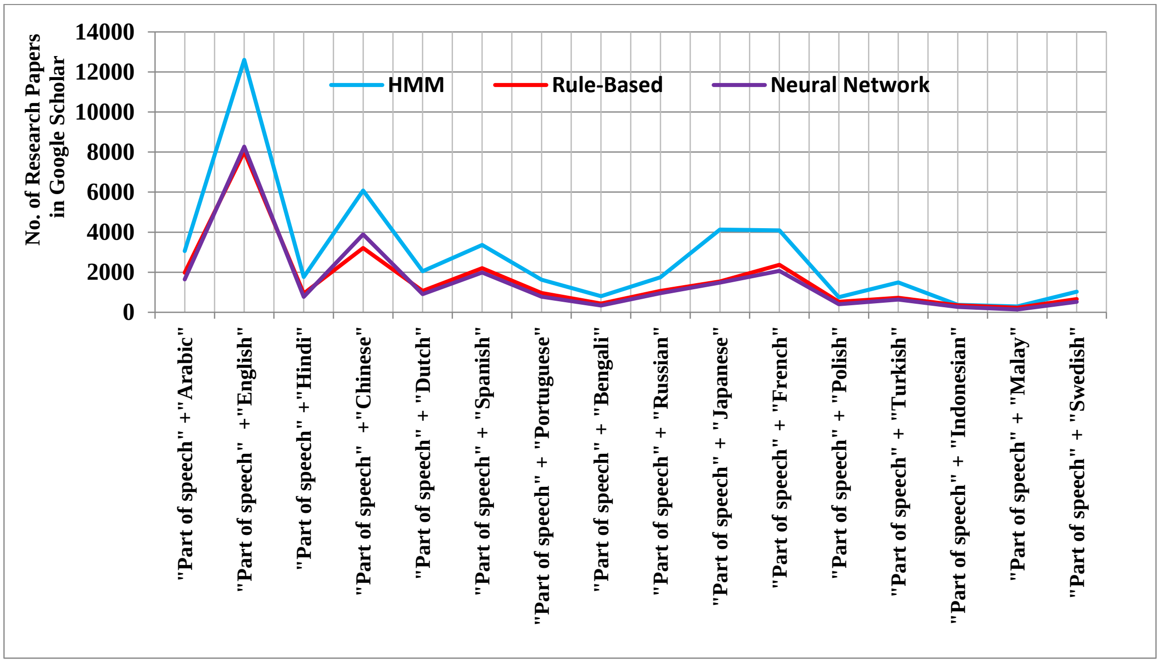


Figure 9: Comparison of different tagger type implementation based different languages.

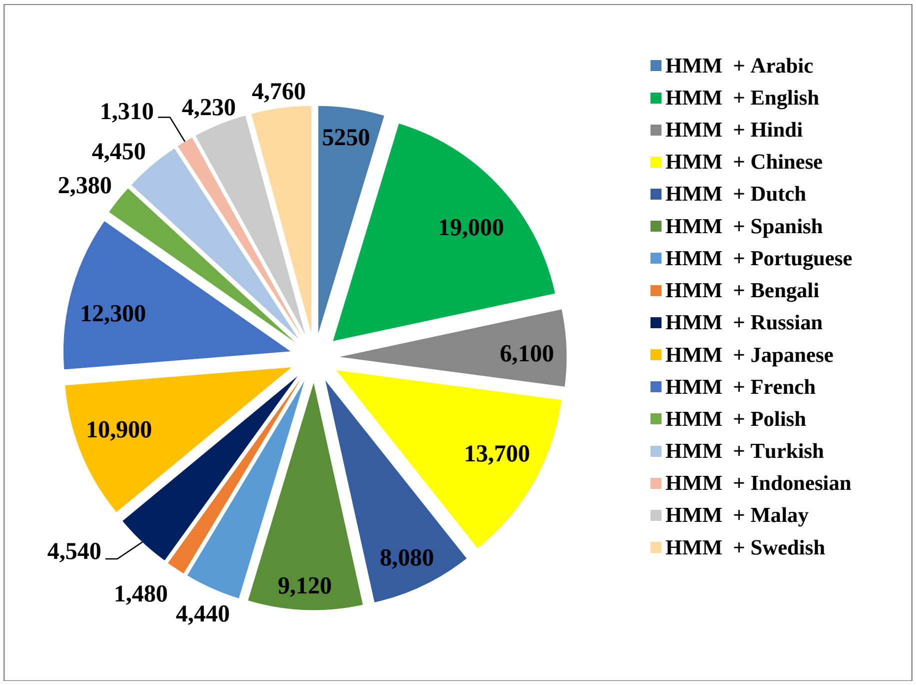


Figure 10: Comparison of different tagger type implementation of HMM based on different languages.

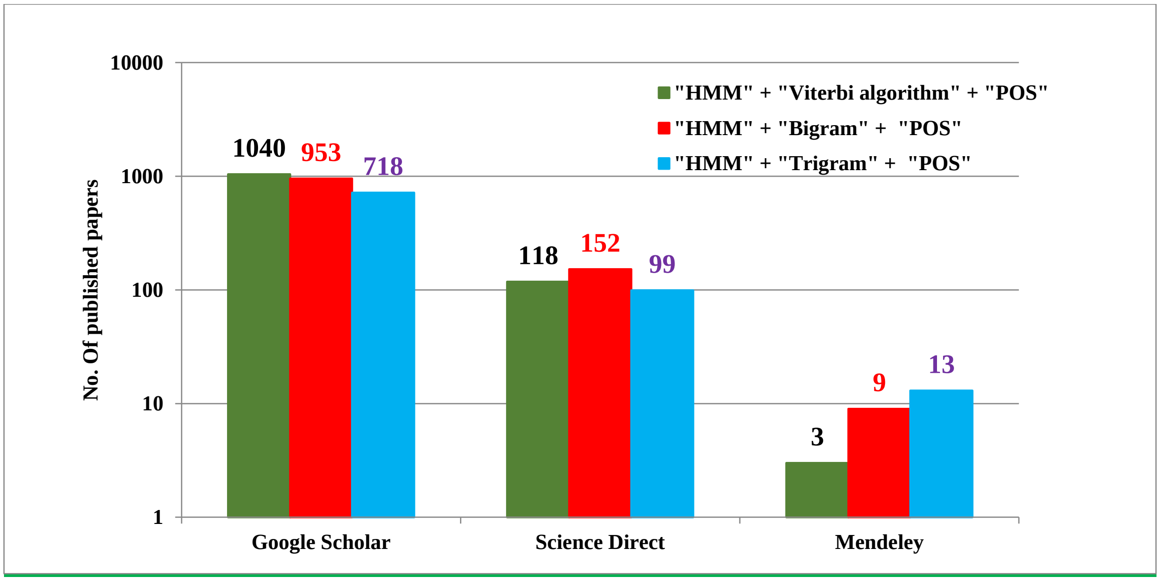


Figure 11: Comparison of number of published papers based on different HMM tagger methods in different scholar databases.

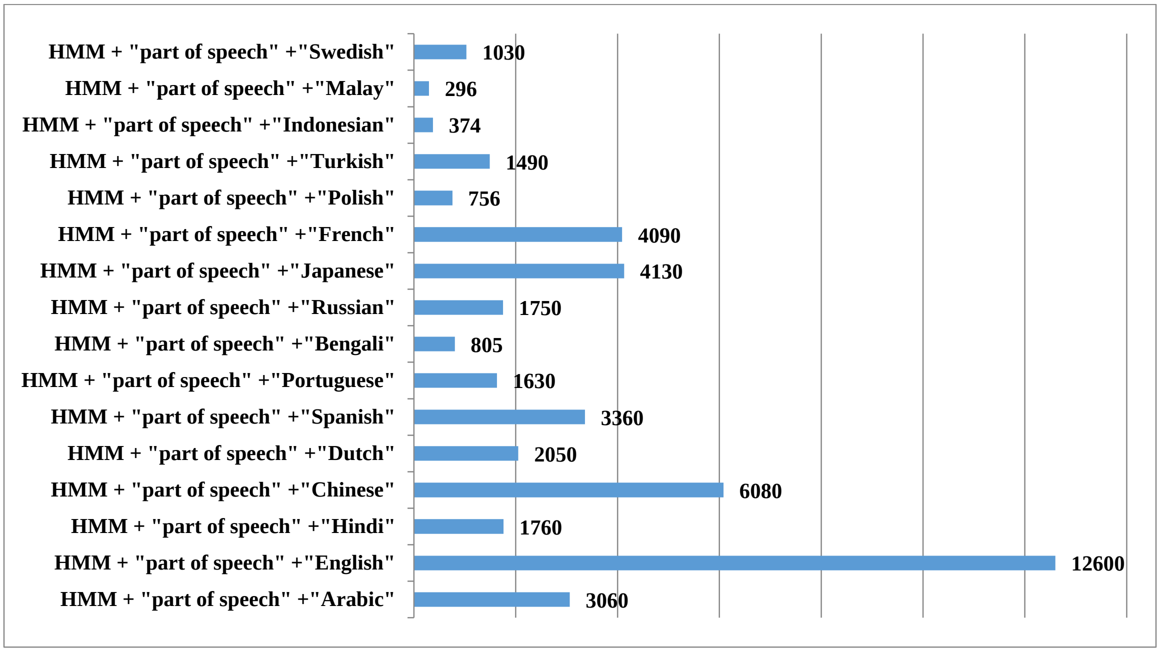


Figure 12: Number of Research Papers on Google Scholar implementing HMM POS tagger Based different types of languages

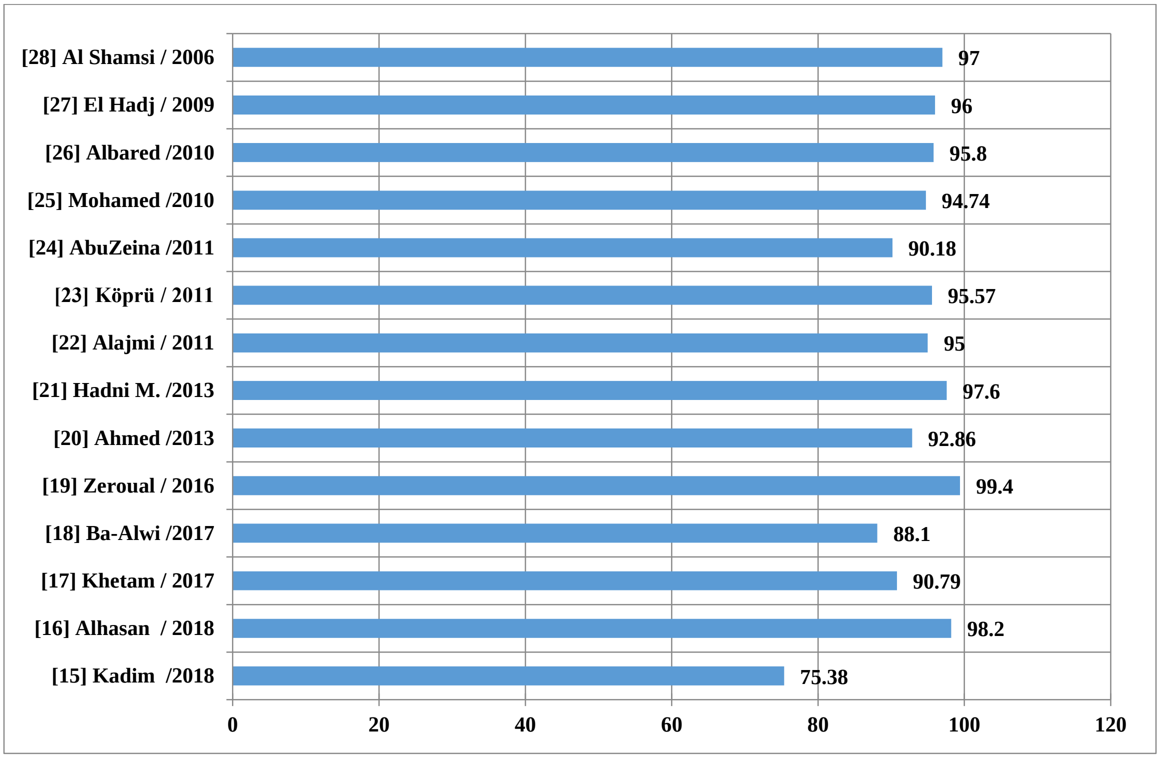


Figure 13: Accuracy of POS Tagging for Different research papers