The Use of Hidden Markov Model in Natural ARABIC Language Processing: a survey

Article *in* Procedia Computer Science · September 2017

DOI: 10.1016/j.procs.2017.08.363

3 authors:

Dima Suleiman
University of Jordan
27 PUBLICATIONS 95 CITATIONS
SEE PROFILE

Arafat Awajan
Princess Sumaya University for Technology
70 PUBLICATIONS 123 CITATIONS
SEE PROFILE

Wael Etaiwi
Princess Sumaya University for Technology
18 PUBLICATIONS 59 CITATIONS
SEE PROFILE

Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 113 (2017) 240–247

Procedia
Computer Science

www.elsevier.com/locate/procedia

The 8th International Conference on Emerging Ubiquitous Systems and Pervasive Networks (EUSPN 2017)

The Use of Hidden Markov Model in Natural ARABIC Language Processing: a survey

Dima Suleiman[a,b,*], Arafat Awajan[a], Wael Al Etaiwi[a]

[a] COMPUTER SCIENCE DEPARTMENT KING HUSSEIN FACULTY OF COMPUTING SCIENCES
PRINCESS SUMAYA UNIVERSITY FOR TECHNOLOGY AMMAN, JORDAN
[b] BUSINESS INFORMATION TECHNOLOGU DEPARTMENT THE UNIVERSITY OF JORDAN AMMAN, JORDAN

Abstract

Hidden Markov Model is an empirical tool that can be used in many applications related to natural language processing. In this paper a comparative study was conducted between different applications in natural Arabic language processing that uses Hidden Markov Model such as morphological analysis, part of speech tagging, text classification, and name entity recognition. Comparative results showed that HMM can be used in different layers of natural language processing, but mainly in pre-processing phase such as: part of speech tagging, morphological analysis and syntactic structure; however in high level applications text classification their use is limited to certain number of researches.

Keywords: **Hidden Markov Model; Arabic Natural Language Processing; Part-of-Speech Tagger; Morphology; Statistical Language Model;Trigram; Bigram, first order logic; second order logic; Text classification; Name Entity Recognition**

* Corresponding author. Tel.: +962795016922
E-mail address: dimah_1999@yahoo.com

1877-0509 © 2017 The Authors. Published by Elsevier B.V.
Peer-review under responsibility of the Conference Program Chairs.
10.1016/j.procs.2017.08.363

1. INTRODUCTION

Natural Language Processing (NLP) applications that utilize statistical approach, has been increased in recent years. One of the most important models of machine learning used for the purpose of processing natural language is Hidden Markov Model (HMM) [1]. Markov Model is a probabilistic model that are considered as sequence classifier such as letters classifier; it calculates the probability of label sequence and chooses the best sequence according to the best possible labels probability distributions. Moreover, Hidden Markov Model is a model that contains a set of state and transitions where transition from one state to another state is determined according to certain input. Each transition contains a value or weight that is determined according to certain probability distribution. Therefore, if certain input causes transmission from state *x* to state *y* then the overall weight will be augmented by the weight *w* that is the value of transition or transition probability between state *x* and state *y*. The probability distribution of a certain transition determines the observation or outcome of a certain state. However, Hidden Markov model is called hidden since the states are not visible and only outcomes can be seen. In our case the input is a sequence of words or letters, so the sequence of words will determine the sequence of states; this sequence represents a chain called Markov chain.

Hidden Markov Model was used in many applications of statistical NLP such as morphological analysis, part of speech tagging (PoST) and text classification. This research provides a comparative study between different applications using Hidden Markov Model in statistical language processing of Arabic language.

The remainder of this paper is structured as follows: Section 2 will explain some Arabic language features. Terminology about Hidden Markov Model, tagging and Markov chain will be introduced in section 3. Section 4 will discuss the related work and finally conclusion will be presented in section 5.

2. ARABIC LANGUAGE FEATURES

Arabic Language has many features that make processing harder than other Languages. Most of Arabic language roots consist of three or four characters, however few have five or more. Arabic language is full of morphology that can be divided into templatic and concatenative. There are no templates for foreign languages words which considered as nonderivative words. Moreover morphemes that are concatenative consist of stem in addition to affixes and clitics. There are three types of affixes: prefix, circumfixes and suffix, also there are two types of clitics: proclitics and enclitics. Generally stem may be preceded by prefix and followed by suffixes; however circumfixes may occurs at the middle of the stem, proclitics at the beginning of the word and enclitic at the end. All these are called morphemes and all morphemes except the stem are optional. The General structure of morphemes can be represented as follows, where a character that represents the optional morpheme is [ ]:

[Proclitic(s)+[Prefix(es)]] + stem + [Suffix(es) + [Enclitic]].

Arabic language also consists of stop or functional words such as pronouns, prepositions, conjunctions and many others, in most of NLP applications functional words were removed in preprocessing phase. Clitics and affixes can be used with stop words, derivative and nonderivative words.

3. TERMINOLOGY

3.1 Hidden Markov Model

Hidden Markov is one of the models that can be used as classifier; Markov consists of a set of state where transition from one state to another depends on certain input. Therefore the transition between states continues until reaching the output state or observation. Moreover, the probability of certain transition depends on the probability of transition from the previous state to current state. The probability model consists of three main elements: experiments with well-defined results, sample space (Ω) which consists of all possible events and finally the event which is subset from the sample space. However HMM depends on conditional probability which states that the probability of occurrence a certain event X depends on the probability of occurrence of previous event Y, this conditional probability can be represented as follows: $p(X|Y)$ which is equal to $P(X \cap Y) / P(Y)$ [1].

Assume that we have three states {Verb, Noun, Adj} as shown in Fig. 1. Transition probability from verb to adj is 0.01, and from adj to verb is 0.02, the transition matrix that displays transition from one state to another can be seen in Table 1.

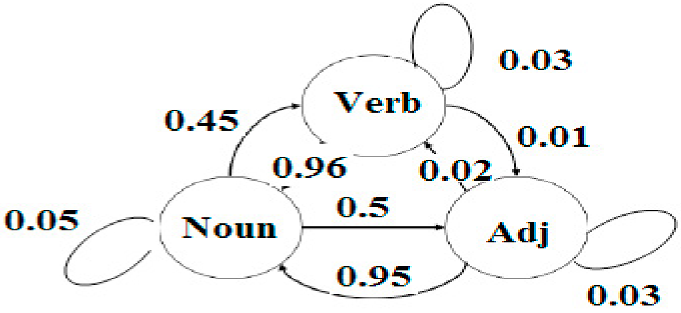

Fig. 1. Markov Model with three states

Table 1. Probability of transition (Transition Matrix)

State		Next State		
		Verb	Noun	Adj
	Verb	P(verb|verb)=0.3	P(Noun |verb)=0.96	P(Adj |verb)=0.01
Previous(current) State	Noun	P(verb| Noun)=0.45	P(Noun | Noun)=0.05	P(Adj | Noun)=0.5
	Adj	P(verb| Adj)=0.02	P(Noun | Adj)=0.95	P(Adj | Adj)=0.3

3.2 The Chain Rule

Chain consists of sequence of words or letters, such as $w_1w_2w_3$... w_n. The probability of $w_1w_2w_3$ …. w_n can be calculated using the following formula: $P(w_1, w_2, w_3,…., w_n) =P(w_1)*P(w_2|w_1)*P(w_3|w_1, w_2)*…*P(w_n \mid w_1, w_2, w_3,….., w_n)$ which is called the probability of chain rule[2]. However by using HMM to calculate the probability of a certain word, there is no need to consider all the previous words from the beginning of the sentence. Thus there are many type of Markov model, such that if the probability of word N depends on the probability of word N-1 this is called first order Markov model or Bigram. Also if the probability of word N depends on the probability of word N-1 and word N-2, then this is called second order Markov model or trigram[3].

3.3 HMM tagging

Tagging or sequence labeling used to map the tag sequence for input sequence. Therefore assume that there are tagging process with the following input: $X_1, X_2, X_3,…., X_m,$ then the output will be a tag sequence or state sequence which is $Y_1, Y_2, Y_3,…., Y_m$. For part of speech tagging, the input will be the sentence and the output will be the tag for each word in a sentence. Thus if we have a sentence with five words the output will be five tags with a certain part of speech. However in machine translation if the input is a sentence in the source language then the label will be a sentence in target language[4].

In order to deal with the tagging problem there are three ways: the first way is rule based method, rule based depends on using already defined rules. However rule based may suffer from several problems such as grammar leak, problem of listing all the rules, and the last problem related to variation of the rules during the time, location and many other circumstances. Also tagging problem can be solved using the statistical based method. Moreover statistical based model depends on statistics and the need of trainable and previously tagged corpus. Finally the last way to deal with the tagging problem is the hybrid model that combines both the rule based and statistical based models together; the last way is most popular and usable approach.

4. RELATED WORKS

In this section a comparative study between different statistical NLP applications has been made. Most of researches that using HMM in their proposed methodology were in morphological analysis and part of speech tagging, however for other applications such as name entity recognition, text classification, machine translation and sentiment analysis, the number of researches decreased.

4.1 Morphological Analysis:

Arabic words either originated from three or four letters which forms the basic block of words. In case of three letters this block is called tri-root and in case of four letters it is called quad-root. Staring from these blocks, other forms of words will be generated with different meanings according to certain morphological rules, these rules are called weight. By detecting the weight of each word we can return the word to its root, however there are more than 300 weights in Arabic to represent all the words. Nevertheless adding clitics and affixes will make the process of finding the root harder[5]. Text classification and processing mainly depends on finding certain features of the word which is not enough to deal with the word itself, thus the segmentation of the word is very important. One of the most significant features of the word is the stem of the word that can be achieved by removing prefixes and suffixes from the word; another representation of the word is to use the word weight.

HMM is used in morphological analysis[6]. Two steps for making analysis were used; the first one consists of dividing of the sentence into words and then finding the morphological units of each word. However the second step is to extract the root of each word. In the first step the words were divided into prefix, suffix and stem, while in the second step the right root was determined by depending on the context. Moreover, in order to determine the right root, HMM was used where the hidden states represented all possible roots, and the observation was used to get the best one. In this step the position of the word into text will be taken into consideration. Also and the high precision was achieved by utilizing already written corpus which consists of 500000 words, this corpus is called NEMLAR Arabic corpus. The correct root was reached by more than 98% of training set.

El Hajar[7] et al. generated part-of-speech tagging using HMM with morphological analysis. Also, HMM was used for representing the structure of Arabic sentences as hierarchal form, the states represent the tags and the syntax of the sentence will determine the transition from one state to another. However the reason for using morphological analysis was to divide the word into prefix, suffix, and stem in order to reduce the size of lexicon. The training of this system used manually tagged corpus.

In order to extract the stem or root of a word, word must be processed also, prefixes and suffixes must be removed from the word, and this can be achieved by a method that depends on HMM[5]. In this model the states represent prefixes, weight and suffixes and transition from one state to another can be done by reading a letter, thus the word is represented by the letters. As a result the highest likelihood of the word will give the best path; moreover the precision of proposed method was 95%. In addition, Markov was used to get the words weight where the surface word will be divided into three parts: prefix at the beginning, weights that the surface word belongs to at the middle and suffixes at the end of the word. The advantage of using weights is that it enables the detection of the word type that may be noun, verb or both. To train the model 15 million words were used.

Morpheme alignment was used to help in stochastic machine translation which is bilingual[8]. In addition to machine translation also segmentation of morpheme was described by using graphical model that is dynamic and unsupervised. They used Hidden Semi-Markov chain that creates special structures to deal with the distribution which is conditional probability; in addition to that they used output nodes that are factored. The target language was used to learn the morpheme and depended on lexical source side, in addition to morpho-syntactic information.

Hidden layers of Semi-Markov were used to represent the morpheme segmentation and observation layer was used to represent sentences which consist of the morphemes in addition to tokens. They used corpus which contains of 6139 phrases that are short from English, Hebrew, and translation of the Bible in Arabic.

4. 2 Part of Speech Tagging.

Part of speech tagging is a process of assigning the appropriate PoST to words which can be determined by depending on a certain context. Many NLP applications depends on PoST as one of preprocessing steps, consequently the performance of PoST tagging will affect the performance of all subsequent preprocessing steps.

The part of speech tagging (PoST) of minimal supervision for Arabic Colloquial Egyptian was studied in[9]. In their research instead of creating specific tools used for dialect, they used unsupervised learning algorithm along with using resources and data that are already exist. The corpus that was used called CallHome Egyptian Colloquial Arabic (ECA) corpus which was already annotated, the baseline tagger that was used is Trigram, Hidden Markov Model (HMM). In order to improve the accuracy from 58.47% to 66.61% the constraint lexicon and affix features were used, in addition, when using joint training the accuracy increased to 68.48%.

New methodology that used Hidden Markov Model (HMM) in (POST) was proposed in[10]. Proposed model used 10 MBs of Arabic corpus. The performance of proposed algorithm that was achieved is 97%, and no transliteration has been used, thus the input of HMM tagger was the original raw text without any conversion. The main reason for using PoST tagging in this work was to extract the Named Entities. Moreover using HMM improved the performance of PoST tagger since it depends on assigning probability values for events in historical events, another reason related to HMM speed in processing large text which is considerable. The tag set that was used is small and rich consists of 55 tag, the small property made the computation of POST tagger feasible, however the richness feature increased the performance and allow good training. The stemmer that was used is Buckwalter and any tagging errors were corrected manually.

Mansour et al.[11] used the same data set used by Habash and Rambow’s, they replaced the morphological analyzer that was used, which were called Hebrew with Buckwalter's. In their research[11] they used the same probabilistic model to enhance Semitics language HMM PoST tagger. This approach increased the accuracy to 96.12% and without missing the advantages of using data driven taggers which was proposed by Diab, thus this model is a hybrid model of Habash, Rambow’s and Diab.

The structure of Arabic sentence is very significant and must be considered in PoST tagging. In their research[12] depended on Arabic sentence structure, where a combination between HMM and morphological analyzer were used. Moreover, Arabic is derivative language, thus morphological analyzer was used to decrease the size of tags lexicon which can be achieved by extracting the stem from the Arabic word. In addition, HMM was used to represent the structure of Arabic sentence in order to consider the sequence of logical linguistics. The states of HMM is the tags and the transition from one state to another was determined according to the sentence syntax. In order to test the new system and to make training, corpus of old text was used, which is a book from third Hijri century that was tagged manually. The experimental results showed that the recognition rate was promising which was 96%.

Yahya et al. [13] used the same technique that was used in[12] which combined HMM with morphological analyzer. They also used the same corpus, book from third century. The tagger was used to annotate and analyze Quran, which is one of traditional Arabic text. The conversion rate was 96%.

Bigram Hidden Markov Model (HMM) was used in[14] to solve the problems of Arabic text tagging. In addition, Viterbi algorithm was used along with using HMM to deal with problem of sparseness. Moreover Viterbi algorithm determines the most suitable tag of the word by computing the probability of the tag. After that, the tag with highest probability will be chosen. However, unknown words PoST were determined by segmenting the word and trying to extract prefix, suffix or linear interpolation of prefix and suffix probabilities. Therefore the experiments showed that the average accuracy was 95.8%., where the training corpus that was used consists of 23146 words.

Corpus containing little training data was analyzed in[15], were hybrid model that augmented the tagger with the weight of morphological analyzer was used. The tagger that was utilized is second-order hidden Markov model that was created specifically to deal with small data. As a result the accuracy of unknown word tag was improved with 96.6% accuracy. Moreover the significant advantage of this research was the dealing with the shortage of Arabic language resources. Nevertheless, this technique may fail to deal with very big training corpus.

The structure of HMM made some limitation in global knowledge such as sentiment analysis. However in[16] a group of researchers proposed a method for using pre-classification HMM in Twitter Part-Of-Speech tagging by using the subjectivity of sentiment analysis. In their approach they generated a general knowledge instead of producing features in details.

An integration between HMM and Rule Based method was used in PoST tagging through a hybrid model[17]. Three POST tags were generated which are, Noun, Verbs and Particles. Two corpora were used for testing and training; they are Holy Quran and Kalimat Corpuses. Kalimat considered as undiacritized Classical Arabic language. The accuracy of this method was 97.6% for Holy Quran and 98% for Kalimat. Corpus of Holy Quran used 33 tags with 77430 words, Kalimat corpus also used 33 tag and both corpuses tags were simplified to three. Other training corpuses can be used to improve the accuracy of determining unknown words tag.

Instead of using three set of tags as in[17], Hadni et al.[18] used four set of tags which are, Noun, Verb, Particle and Quranic. [16] Used the same hybrid model used in[17] which combine HMM with Taani's Rule-based approaches. In their experiments they used Holy Quran Corpus which was undiacritized and the accuracy of proposed method is 97.6%. Moreover, they will try to improve the unknown words tagging and to apply their model in Multi-Word expressions.

Aliwy proposed a new approach for PoST[19], his approach based in combining different taggers, he used a master-slaves technique where the master is HMM and slaves are any other taggers. In their work, the techniques that were used as slaves were maximum match (MM) and Brill taggers. The technique start with tagging the sentence using slaves, the resulted tags will be used to change the probability of the master tagger tags for the same sentence. The experiments were made using 45k words in private Arabic corpus. Also, Brown corpus was used for English language. The selection of the best tag that was proposed by slave's tagger was controlled by master tagger such as HMM.

Bidirectional HMM-based approach[20] was called bidirectional. However bidirectional approach called like this, since it considered both the future and the past element of a certain word in order to determine its morph-syntactic state. In their research they combined both reverse and direct taggers in both senses in order to tag the same words sequence using HMM which depends on past elements. The cost of implementing and programming two different taggers was low due to using the same resource and make changes only in training process, in their experiments they used Nemlar corpus, which consists of 500,000 words.

The main problem of using small amount of data for training process was the determining of the tag of unknown words. However, this problem became significant especially when dealing with languages that have a huge vocabularies and complex morphology like Arabic. In order to solve the previous problem[21] proposed an approach that used second order Hidden Markov Model (trigram). In addition, they also studied the performance of taggers, and how the performance can be affected if they combined with morphological analyzer which depends in lexicon. In their work they also implemented, evaluated and empirically defined different lexical models. The training corpus that was used consists of 24 tag set with 29300 words. In the future they suggested to increase the size of training corpus and to map unknown word to the pattern of known word in order to improve pattern identification process.

4.3 Text Classification

The existence of large amount of information recently, results in various problems such as information filtering and extraction, automatic metadata generation and semantic indexing of the documents, all these problems can be solved using text classification[22]. However, text classification is a technique used to sort the documents according to certain category. In addition, text classifying is very interested since manual text categorization is expensive and it is also a time consuming process. Therefore, text classification can be used in many applications such that mail routing, attribution of authorship, monitoring of news and indexing of scientific articles automatically and many others. Text processing can be classified as information retrieval and extraction, segmentation of text and tracking of events. However text categorization can be achieved by using Hidden Markov Model as proposed in[23].

Arabic text classification was developed using statistical learning in 2006[22], they used the stem of Arabic words; also they extracted the features vector structure of the stem. In addition, in their research the semantic of the document was used to create systems for question and answer. However, the components of features vector are divided into topics their numbers equal to number of topics in corpus, since the corpus will be divided into topics. The posterior value of each feature component will be equal to the ratio between how may time the stem appears in a topic, to the total numbers of words in that topic, in order to make experiments the corpus that was used consists of 30 topics, and each topic consists of 15 fatwas, three of them selected randomly for training to end up with 450 fatwas.

Hidden Markov Model was used to extract features in text categorization process such as root and weights of morphology [24]. While stem and weight are very close to each other, the weight in some prefixes will remain without being removed. After computing the weights, it transfers the weight into a form that is unified; this can be done by grouping different states (plural, single, past, future) that have the same meaning. Each state in HMM will represent a letter of the word and the word can be formed by many transitions within the states. In their research they produced parameters of Hidden Markov Model by a set of steps. In first step Arabic dictionary was used to collect the pattern of the word's root, then words with different forms were found by using morphological rules, and finally the final results were utilized in order to train the model after adding suffixes and prefixes of different forms.

N-gram as a feature can be used by HMM classifier to classify Arabic documents[25, 26]. However, the main feature that combines similar text together is in term of themes[25]. Therefore the documents that have the same theme will be combined in the same group, theme can be modeled using HMM with term as observations and compound term as state. Consequently, experimental results showed that the HMM is very convenience in text classifications. On the other hand, Arabic poetry used HMM in order to determine the Authorship Attribution[27] where inputs for HMM were length of words, rhyme, first word character and the length of the sentences, experiments was conducted in text wrote by 33 Arabic poet the accuracy was 96.97%.

5. CONCLUSIONS

Natural Language Processing is one of the most brilliant topics in computer science. Processing of natural language can occur either by using rule base approach, statistical approach or hybrid of both. However, the most common model that can be used for statistical NLP is HMM. In this research, a comparative study is made between different applications that use HMM in their processing of Arabic language text. Moreover, the study concluded that HMM can be used in different layers of NLP, but the significant affect was in Morphological Analysis and part of speech tagging that is used in most of NLP applications in preprocessing phase. On the other hand limited number of high level NLP application can use HMM such as Text classifications.

REFERENCES

1. Lussier E., Markov Models and Hidden Markov Models_ A Brief Tutorial, INTERNATIONAL COMPUTER SCIENCE INSTITUTE, 1998
2. Jurafsky D., Martin H, Speech and Language Processing. Copyright c 2016. All rights reserved. Draft of November 7, 2016.
3. Al-Anziand F., AbuZeina D., A SURVEY OF MARKOV CHAIN MODELS IN LINGUISTICS APPLICATIONS, 10.5121/csit.2016.61305
4. Christopher D. Manning , Hinrich Schütze, Foundations of statistical natural language processing, MIT Press, Cambridge, MA, 1999.
5. Alajmi A. F., Saad E. M. and Awadalla M. H., Hidden markov model based Arabic morphological Analyzer, International Journal of Computer Engineering Research Vol. 2(2), pp. 28-33, March 2011
6. Boudlal A., Belahbib R., Lakhouaja A., Mazroui A., Meziane A., Bebah M., A Markovian Approach for Arabic Root Extraction, The International Arab Journal of Information Technology, Vol. 8, No. 1, January 2011
7. El-Hajar A, Hajar M, Zreik K (2010). A System for Evaluation of Arabic Root Extraction Methods. fifth international Conference on Internet and Web Applications and Services.
8. Naradowsky J. , Toutanova K., Unsupervised bilingual morpheme segmentation and alignment with context-rich hidden semi-Markov models, Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, June 19-24, 2011, Portland, Oregon
9. Duh K., Kirchhoff k. POST tagging of dialectal Arabic: a minimally supervised approach, In Proceedings of the acl workshop on computational approaches to semitic languages, pp. 55-62. Association for Computational Linguistics, 2005.
10. Al Shamsi F., Guessoum A., A Hidden Markov Model –Based POST Tagger for Arabic, Journées internationales d'Analyse statistique des Données Textuelles, 2006
11. Mansour S., Sima'an K., Winter Y., Smoothing a Lexicon-based POST Tagger for Arabic and Hebrew, Proceedings of the 5th Workshop on Important Unresolved Matters, pages 97–103, Prague, Czech Republic, June 2007. c 2007 Association for Computational Linguistics
12. ElHadj, Y.O.M., I.A. AlSughayeir, A.M. Khorsi, A.M. Alansari, 2009. Morphology analysis of the Holy Quran: An indexed Quran text database (in Arabic). Proceeding of the 5th International Conference on Computer Sciences Practice in Arabic, Rabat, Morocco, May 2009
13. Yahya O. Mohamed Elhadj(2009)," Statistical Part-of-Speech Tagger for Traditional Arabic Texts", Journal of Computer Science 5 (11): 794-800.
14. Albared, M., Omar, N., Ab Aziz, M.J., and Nazri, M.2010. Automatic part of speech tagging for Arabic: An experiment using bigram hidden markov model. Lecture Notes Comput. Sci. Springer, 6401: 361-370. DOI: 10.1007/978-3-642-16248-0_52.
15. Albared M., Omar N., Juzaiddin Ab AzizMohd. Improving Arabic Part-of-Speech Tagging through Morphological Analysis,2011
16. Sun S., Liu H., Lin H., Abraham A., Twitter Part-Of-Speech Tagging Using Pre-classification Hidden Markov Model, Systems, Man, and Cybernetics (SMC), 2012 IEEE International Conference on, 2012.
17. Hadni M., Ouatik S. E., Lachkar A., and Meknassi M., Hybrid Part-Of-Speech Tagger for Non-Vocalized Arabic Text. International Journal on Natural Language Computing (IJNLC) Vol 2 (2013).
18. Hadni M., Ouatik S., Lachkar A., and Meknassi M., IMPROVING RULE-BASED METHOD FOR ARABICPOST TAGGING USING HMM,2014.
19. Aliwy A., Combining POST Taggers in Master-Slaves Technique for highly inflected languages as Arabic, IEEE,2015
20. Kadim A., Lazrek A., Bidirectional HMM-based Arabic POST tagging, 0.1007/s10772-015-9303-7, 2015
21. Albared M., Al-Moslmi T., Omar N., Al-Shabi A., Mutaher Ba-Alwi F., PROBABILISTIC ARABIC PART OF SPEECH TAGGER WITH UNKNOWN WORDS HANDLING, Journal of Theoretical and Applied Information Technology, 31st August 2016. Vol.90. No.2
22. Reda A. El-Khoribi and Mahmoud A. Ismael. "An Intelligent System Based on Statistical Learning For Searching in Arabic Text". In the ICGST International Journal on Artificial Intelligence and Machine Learning, AIML, 6(3). Pages 41-47. 2006
23. Yi K., TEXT CLASSIFICATION USING A HIDDEN MARKOV MODEL, A thesis submitted to the Faculty of Graduate Studies and Research in partial fulfillment of the requirements for the degree ofDoctor ofPhilosophy,2005
24. Alajmi A. F., Saad E. M, Awadalla M H, DACS Dewey index-based Arabic Document Categorization System, International Journal of Computer Applications (0975 – 8887) Volume 47– No.23, June 2012
25. Kechaou Z., Kanoun S., A new-arabic-text classification system using a Hidden Markov Model,International Journal of Knowledge-based and Intelligent Engineering Systems 18 (2014) 201–210, DOI 10.3233/KES-140297
26. Graovac J., Text Categorization Using n-Gram Based Language Independent Technique
27. Ahmed A. Mohamed R. BellafkihMostafa, Mohammed A., Authorship Attribution in Arabic Poetry Context Using Markov Chain classifier. IEEE, 2015