Journal of Computer Science 5 (11): 794-800, 2009
ISSN 1549-3636

Statistical Part-of-Speech Tagger for Traditional Arabic Texts

Yahya O. Mohamed Elhadj
Department of Computer Science, College of Computer and Information Sciences,
Imam Muhammad Bin Saud University, P.O. Box 8488, Riyadh 11681, KSA

Abstract: Problem statement: This study presented the development of an Arabic part-of-speech tagger that can be used for analyzing and annotating traditional Arabic texts, especially the Quran text. **Approach:** It is a part of a project related to the computerization of the Holy Quran. One of the main objectives in this project was to build a textual corpus of the Holy Quran. **Results:** Since an appropriate textual version of the Holy Quran was prepared and morphologically analyzed in other stages of this project, we focused in this work on its annotation by developing and using an appropriate tagger. The developed tagger employed an approach that combines 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. On another hand, HMM is used to represent the Arabic sentence structure in order to take into account the linguistic combinations. For these purposes, an appropriate tagging system has been proposed to represent the main Arabic part of speech in a hierarchical manner allowing an easy expansion whenever it is needed. Each tag in this system is used to represent a possible state of the HMM and the transitions between tags (states) are governed by the syntax of the sentence. A corpus of some traditional texts, extracted from Books of third century (Hijri), is manually morphologically analyzed and tagged using our developed tagset. **Conclusion/Recommendations:** It is then used for training and testing this model. Experiments conducted on this dataset gave a recognition rate of about 96% and thus are very promising compared to the data size tagged till now and used in the training. Since our Holy Quran corpus is still under revision, we did not make significant experiments on it. However, preliminary tests conducted on the seven verses of AL-Fatiha showed an encouraging accuracy rate.

Key words: Hidden Markov models, Arabic morphological analysis, holy Quran, text corpus, classical Arabic, modern standard Arabic

INTRODUCTION

Part-Of-Speech tagging (POS tagging) is the process by which a specific tag is assigned to each word of a sentence to indicate the function of that word in the specific context[1].

POS tagging is considered as one of the basic tools and components necessary for any robust Natural Language Processing infrastructure of a given language[2]. It is needed in many fields of linguistic processing, starting from the simpler ones as text phrasing and alignment, to the more elaborate ones as syntax and semantic analysis and ending up with linguistic processes that is heavy as machine translation. Moreover, POS tagging is also considered as first stage for analyzing and annotating corpora[3].

Many taggers have been developed for different languages and used to build various kinds of applications. These applications include speech synthesis, natural language parsing, information retrieval and information extraction.

This study is a part of a project aiming to develop a computerized-environment of the Holy Quran. One of the main tracks in this project was to build an annotated textual corpus of the Holy Quran that can be used as a reference for different linguistic studies related to the Holy Quran. The similarity between Quranic terms is a good example of these studies and is being treated now by our team[4].

We have started, in this project, by preparing an authentic fully diacritised textual version of the Holy Quran since such one was not available for the use in the domain of research, as our knowledge goes. We have then focused on a manual morphological analysis of this version[5]. Each word in the Holy Quran is then split into four parts: Prefixes, stem, root and suffixes and stored in an indexed database. Words are kept in their original context (Quranic verses). It is worth to mention that we are aware of the existence of some available Arabic morphological analyzers, such as Sakhr (www.sakhr.com/), RDI (www.rdi-eg.com), Buckwalter[6]. However, due to the specificity of the Holy Quran and the complexity of its linguistic structure, we preferred to perform a manual morphological analysis.

Having passed the morphological stage, we want to start the tagging process of the parts of speech related to the Holy Quran. Contrary to the morphological analysis task where a manual work was considered more appropriate, we think that using a very well designed POS tagger may be benefit able and can accelerate this annotation task. However, this tagger should be built based on the true Arabic grammar and trained on an enough corpus of texts where correct and diversified linguistic structures are employed. The tagger may be used to annotate classical Arabic texts (النصوص التراثية القديمة). In the case of the Holy Quran, it can be at lest used as a first pass and if necessary followed by a manual revision to ensure the required high quality of precision.

As far as we know, very few works have been interested, in the literature, to the POS tagging of the Holy Quran. A team at Haifa University is working on a computational system for morphological analysis and annotation of the Quran for research and teaching purposes[7,8]. Their studies indicate that a set of finite-state based rules describing the morpho-phonological and morpho-syntactic phenomena of the Quranic language have been developed. A phonemic transcription of the Quran text is performed using ASCII notations and is then divided into three classes: Closed-class words, nominal bases and verbal bases. However, although this approach is interesting, we think that it may not be so accurate as required for the Holy Quran.

In this study, we present the development of a part of speech tagger based on the Arabic sentence structure. It combines morphological analysis approach with statistical one based-on Hidden Markov Model (HMM). For this reason, an appropriate tagging system has been firstly proposed and used to represent the main Arabic parts of speech. A baseline of this study with preliminaries results was presented in the conference medar'09 held at Cairo, Egypt[9].

Once such a tegger is built and trained it can be progressively used to annotate our developed Holy Quran corpus. Each time a chapter (part) of the Holy Quran is tagged and verified it can be used to retrain the model again in order to integrate new knowledge related to the linguistic structure of the Holy Quran. It is known that the Holy Quran is composed of different chapters (114 surahs) with variable lengths. So, we can start using the model on short-to-medium chapters to reduce the overhead of manual verification. With this process, we think that the model becomes quickly more and more accurate. However, we expect that a phase of manual verification still needed to ensure high precision required for the Holy Quran.

Pos-tagging techniques: Arabic POS tagging (APOS tagging) is not an easy task due to the high ambiguity results from the absence of diacritics and also from the complexity of the Arabic morphology. Consider the following example: "عالم علم رجلا" (a scholar taught a man). Each word in the above example has more than one morphological analysis. The APOS tagger is responsible for assigning to each word the most appropriate morphological tag.

There are three general approaches to deal with the tagging problem:

Rule-based approach: consists of developing a knowledge base of rules written by linguists to define precisely how and where to assign the various POS tags.

Statistical approach: consists of building a trainable model and to use previously-tagged corpus to estimate its parameters. Once this is done, the model can be used to automatically tagging other texts. Successful statistical taggers were built during the last years and are mainly based on Hidden Markov Models (HMMs).

Hybrid approach: Consists in combining rule-based approach with a statistical one. Most of the recent study uses this approach as it gives better results.

Different Arabic taggers have recently emerged, some of them are developed by companies (Xerox, Sakhr, RDI) as commercial products, while others are a result of research efforts in the scientific community[10-15]. Among these studies, khoja[10] combines statistical and rule-based techniques and uses a tagset of 131 basically derived from the BNC English tagset. Freeman's tagger[11] is based on the Brill tagger and uses a machine learning approach. A tagset of 146 tags, based on that of Brown corpus for English, is used. Maamouri and Cieri[12] base their research on the automatic annotation output produced by the morphological analyzer of Tim Buckwalter[6]; it achieved an accuracy of 96%. Diab *et al.*[13] use Support Vector Machine (SVM) method and the LDC's POS tagset, which consists of 24 tags. Banko and Moore[14] 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 and achieves an accuracy of about 96%. Tlili-Guiassa[15] uses a hybrid method of based-rules and a memory-based learning method. A tagset composed of symbols from Khoja's tagger and new ones is used and a performance of 85% was reported.

Almost all of these taggers, either use tagsets derived from English which is not appropriate for Arabic, either they rely on a transliteration of the Arabic input text. Moreover, these taggers are generally developed for what we call Modern Standard Arabic (MSA) and thus may not appropriate for the Classical Arabic which is the language of the Holy Quran. For this reason, we preferred to develop our own tagger that relies on a correct Arabic sentence structure. Notice that, in our knowledge, the structure of the Arabic sentence was not generally taken into account during the tagging process and few works are interested to that[16].

As we will explain, an appropriate tagging system is firstly proposed to represent the main Arabic part of speech and then a HMM-based statistical approach combined with morphological analysis is considered.

METHODS AND MATERIALS

Our approach for Arabic pos-tagging

In this study, a form of combination between statistical and linguistic approaches will be employed, so that the processing will be performed in two levels. In the first level, text is firstly normalized and tokenized into words and then morphologically analyzed. The morphological analysis is used as input module to reduce the size of the needed tags' lexicon by segmenting Arabic words in their prefixes, stems and suffixes. This is very important due to the fact that Arabic is a derivational language. For this purpose, an appropriate tagging system has been proposed to represent the main Arabic part of speech in a hierarchical manner allowing an easy expansion whenever it is needed.

In the second level, an appropriate statistical model based on the internal structure of the Arabic sentence is used to recognize the morphological characteristics of the words for the entered text. The use of the linguistic internal structure of the Arabic sentence will allow us to identify logical sequences of words and consequently their corresponding tags. Since the probability of a certain word (or its tag) occurrence depends on the words preceding it in a given context, the HMM will be the best suitable statistical model to keep track of this history.

A linguistic study is conducted to determine the Arabic sentence structure by identifying the different main forms of both nominal and verbal sentences. Based on that, a HMM model is then used to represent this structure. 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. Transition' probabilities are calculated using a smoothed tri-gram and a special processing is used to handle unknown words to determine their lexical probabilities.

Before giving the details of our Arabic POS tagger, a linguistic study of Arabic words and grammatical structures will be required for the purpose of coding morphological characteristics and for extracting the most appropriate structure for common Arabic sentence' forms.

Description of the tagging system: We investigated the principle aspects of Arabic morphology and grammar. The following is a brief review of those aspects. The Arabic verbal structures are composed of three classes: noun (اسم), verb (فِعل) and that we will call particle (حَرف).

Noun: It is either a name or a word that describes a person, thing or idea. It could be definite or indefinite and can be subcategorized by the person (narrator, interlocutor and absent), number (Singular, Dual, Plural), gender (Masculine, Feminine) and grammatical cases (nominative "الرفع", accusative "النصب", genitive "الجر"). Figure 1 gives a main classification of the noun and its prominent ramifications.

Verb: It is a word that denotes an action and could be combined with some particles. In term of tense (Fig. 2), the verb could be past (perfect), present (imperfect) or imperative. A future verb tense exists, but it is a derivative of the present tense that you achieve by attaching a prefix to the present tense of the verb.

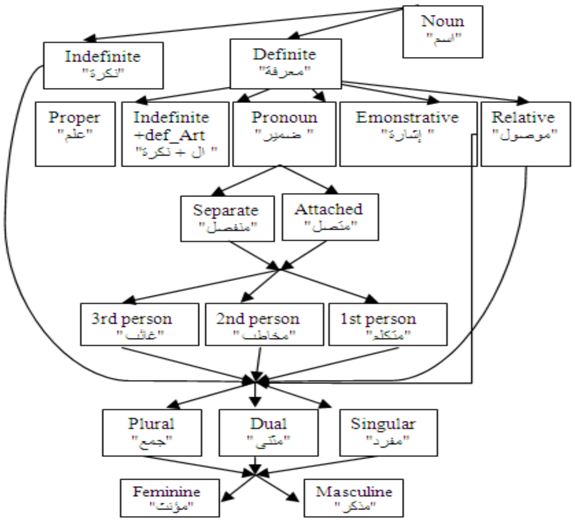

Fig. 1: Noun and its sub-categories

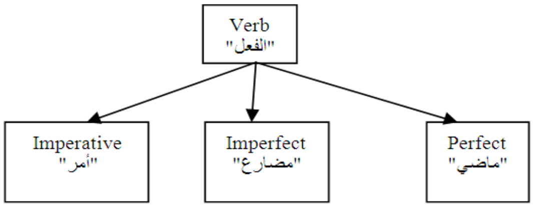

Fig. 2: Verb and its temporal-forms

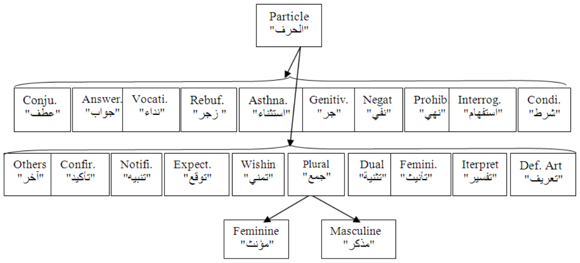

Fig. 3: Main groups of particles

Particles can be added as prefixes and/or suffixes indicating the number, gender and person of the subject, like for example: يقول (he says), قالت (she said), يقولان، تقولان، يقولون (they say). Three moods are possible for verbs: Indicative "الرفع", subjunctive "النصب" and jussive "الجزم".

Particle: This class includes everything that is neither a verb nor a noun. It contains for example, genitive letters "حروف الجر", prepositions of coordination, conjunction as well as the functional words like " إن وأخواتها، كان وأخواتها" which influence the upcoming words analysis. Fig. 3 gives an example of the classification of particles according to their functions.

Proposed tagset: The previous classification is used to develop an appropriate tagging scheme. The hierarchy of parts of speech allows the development of meaningful tags easily expandable to include more details and precision about the Arabic units whenever it is needed.

As we have seen before, the noun could be defined or undefined. We will give the noun in its generic format the symbol "No". In its defined format, it will get the symbol "NoPr" if it is a proper name, the symbol "NoPn" if it is a pronoun, "NoDe" if it is a demonstrative pronoun, "NoRe" if it is relative pronoun. The pronoun could be attached-to "متصل" or separated-from "منفصل" the next word. So we will use "NoPnAt" to tag the first one and "NoPnSe" to tag the second one. To indicate the gender (masculine, feminine), number (singular, dual, plural) and person (1st, 2nd, 3rd), we will add respectively the letters M, F, S, D, P and numbers 1, 2 or 3.

As far as the verb is concerned, it will be given the symbol "Ve" globally. It takes "VePe" for the Perfect, "VeIf" for the Imperfect and "VeIa" for the imperative.

Regarding the class of particles, tags are specified only for some ones that are of subject matter for our research in its current phase. Among those, "PaDe" is used to tag the identifier (أل), "PaDu" and "PaPl" are respectively used for tagging particles indicating the number (dual and plural). For indicating the gender, the letters M or F can be used. The remaining particles are assigned the tag "PaOt", but they can be tagged separately following the same logic. "Pa" is the global tag given to the particle if we do not need to distinguish a particular one.

Finally, we will assign to the punctuation signs (., ?, !,) the symbol "Pu". The digits and dates are denoted by the symbol "Nu". Notice that in the Holy Quran and the Classical Arabic, punctuation marks are not generally employed.

Specification of the sentence structure (model architecture): A linguistic study has been conducted to extract common types of formulations of the Arabic sentence, so that it can serve as architecture of the statistical model. The references of this study were the old morphology books and modern studies concerned with sentence structures in the Arabic language such as Harkat, Mutawakkil, Al-Rahhali, Al-Shukri, Yaqut.

The sentence in the Arabic language is either nominal like in "الشمس ساطعة" (the sun is bright) or verbal like in "يلعب الأطفال الكرة" (the children play the ball). Each of them may have different forms and styles. A list of more than 100 ways of common grammatical structures in the Arabic language has been surveyed. It covers the general syntactical analysis and detailed morphological analysis of the nouns and verbs.

Structure of nominal sentences: Different forms of formulation have been identified for nominal sentences. They can be represented by the graph (Fig. 4) in terms of sequences, where V, N and P respectively denote Verb, Noun and Particle. S and E are special states, used to represent the start and the end of the nominal phrase. Notice that a loop on a state indicates certain number of repetitions of this symbol and an arrow between two sates, means that first one may be followed by the second one depending on its direction.

For the sake of the simplicity, we preferred to represent the nominal sentence in this compacted form. It can be expended by replacing N (noun), V (verb) and P (particle) by their subcategories hierarchically until terminal symbols are reached. This also is the same for the following verbal sentence structure, where a short form is used.

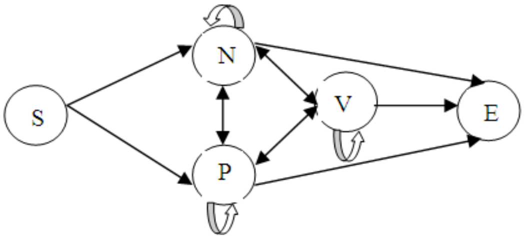

Fig. 4: Structure of nominal sentences

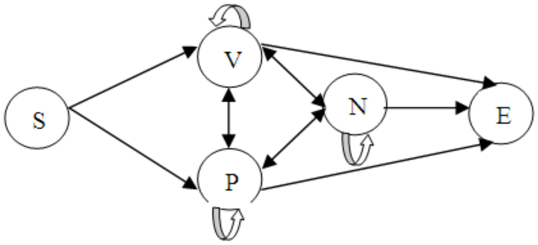

Fig. 5: Structure of verbal sentences

Structure of verbal sentences: Verbal sentence structure can be represented by a graph as in the Fig. 5. This means that a verbal sentence starts either by a verb or a particle and is fallowed by any combination of the main parts of speech.

Architecture of the statistical model: Although the previous representation of both nominal and verbal sentence structures can be seen as trivial and straightforward, they are very interesting for specifying the architecture of our HMM model. It consists to combine them in a one graph and to replace each state by the underlying parts of speech and then expand them to include their subcategories as we have specified in the description of the tagset. Each state in the new graph will then represent a valid tag from our lexicon. So, this graph can be used as the architecture of the HMM model. Determination of the model parameters will be discussed in the following paragraphs.

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

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

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

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

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

$$P(T = t_1t_2\cdots t_n) = \prod_{i=1}^{n} P(t_i \mid t_{i-2}t_{i-1})$$

A tagged training corpus is used to compute $P(t_i \mid t_{i-2}t_{i-1})$, by calculating frequencies of trigrams and bigrams (respectively $f(t_{i-2}t_{i-1}t_i)$ and $f(t_{i-2}t_{i-1})$) as follows**:**

$$P(t_i \mid t_{i-2}t_{i-1}) = f(t_{i-2}t_{i-1}t_i) / f(t_{i-2}t_{i-1})$$

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[17]:

$$\lambda_1 * P(t_i \mid t_{i-2}t_{i-1}) + \lambda_2 * P(t_i \mid t_{i-1}) + \lambda_3 * P(t_i)$$

where, $\lambda_1 + \lambda_2 + \lambda_3 = 1$.

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

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

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

$$P(w_i \mid t_i) = f(w_i, t_i) / f(t_i)$$

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 tags given a sequence of input words. The Viterbi algorithm is used to perform this task.

RESULTS AND DISCUSSION

Corpus preparation: We remember that our ultimate goal is to build an Arabic POS tagger that can be used to annotate our morphologically analyzed Holy Quran text. We aim also that the tagger be able to work on old traditional texts "نصوص تراثية قديمة". For this reason and since the Holy Quran corpus is still under a final verification, we have preferred to enlarge our small corpus, previously prepared, which is composed of some texts extracted from ALJAHEZ's book entitled "Albayan-wa-tabyin" (255 Hijri). It is obtained from "Ashamila" library, which is downloadable from the link: http://www.shamela.ws. Although these texts are not so old (third century Hijri), their styles may vary greatly from those of MSA employed nowadays. A manual morphological analysis followed by a POS tagging of a new part of this corpus has been performed. With the previous prepared part, the ready corpus counts 56312 total tokens (generated by morphological analysis) with a 6439 unique ones ranged in 2000 sentences. Among these counts, there are 25269 nouns, 6825 verbs and 22218 particles. The remaining tokens (2000) represent sentences boundaries (start symbol denoted by <s>). Another set composed of 300 sentences containing 15892 total tokens is also prepared to be used as a basis of testing.

Data-sets and evaluation: Our model is trained on the large part of the corpus described above and tagged using 13 tags: 3 subcategories of verbs, 6 subcategories of nouns and 4 subcategories of particles. It is tested on different subsets of the second part of the corpus, which contains about 16000 tokens. To evaluate its performance, we have used the F-measure defined as follows: (2*P*R*)/(P+R), where P and R denotes precision and Recall respectively. They are calculated using the total number of correct assigned tags (Nc), total number of assigned tags (Na) and the total number of the assigned tags in the test-set (Nt): P = Nc and R = Nc/Nt.

We have obtained an accuracy of about 96%, which is very encouraging compared to the size of the tagset used till now.

To evaluate the model on the Holy Quran, preliminaries tests have been conducted on the seven verses of Al-Fatiha. The result shows an accuracy rate of almost 94% (Table 1).

Table 1: Error rates on the seven verses of Al-Fatiha

Verse number	No. of words	Error rate (%)
1	7	0
2	7	0
3	4	0
4	4	0
5	5	0
6	6	0
7	15	3/15 = 20 %
Global error	48	3/48 = 6.25%

We will start using the model on the whole text of the Holy Quran as soon as its final verification terminates. As we said in the introduction, the model will be used progressively in order to integrate linguistic knowledge of the Holy Quran in the model and then increase its efficiency.

CONCLUSION

In this study we have presented an Arabic Part-Of-Speech tagger that uses a HMM model to represent the internal linguistic structure of the Arabic sentence. We have conducted a linguistic study to determine the main Arabic Part Of Speech and to specify different common forms of Arabic sentences. After that, an appropriate tagging system has been proposed to represent these main Arabic parts of speech in a hierarchical manner allowing an easy expansion whenever it is needed. Next, a suitable architecture of the HMM model was specified based-on the structure of both nominal and verbal sentence. Having done this, a corpus composed of traditional texts extracted from books of third century Hijri was created. Parts of it were manually tagged and used to train and to test the tagger. Performance evaluation has shown an accuracy of about 96%. However, although this represents a very good result compared to the size of the training corpus, we have to increase our tagged corpus and to conduct further tests on more interesting dataset to evaluate the real performance of this approach. Moreover, preliminary tests on some verses of the Holy Quran show an acceptable level of precision. This precision is expected to increase when some parts of the Holy Quran are used in the training.

In addition to the annotation of the Holy Quran, we plan to use the developed tagger for other research activities in a variety of ways, especially for applications dealing with traditional texts "النصوص التراثية".

ACKNOWLEDGMENT

This study is supported by King Abdulaziz City for Science and Technology under the grant number AT-

25-113, Riyadh, Saudi Arabia. Many thanks for all the team members of the Holy Quran Project.

REFERENCES

1. Jurafsky, D. and J.H. Martin, 2008. Speech and Language Processing: An Introduction to Speech Recognition, Computational Linguistics and Natural Language Processing. 2nd Edn., Prentice Hall, ISBN: 10: 0131873210, pp: 1024.
2. Atwell, E., L. Al-Sulaiti, S. Al-Osaimi and B. Abu-Shawar, 2004. A review of Arabic corpus analysis tools. Proc. JEP-TALN'04 Arabic Language Processing.
3. Alansary, S., M. Nagi and N. Adly, 2008. Towards analyzing the International Corpus of Arabic (ICA). Proceeding of the 8th International Conference on Language Engineering, Egypt.
4. AlSughayeir, I.A., A.M. Khorsi, A.M. AlAnsari and Y. AlOhali, 2009. Search engine for the similarity in the terms of the Holy Quran (in Arabic). Proceeding of the International Conference on the Glorious Quran and Contemporary Technologies, Oct 2009, King Fahd Complex for the Printing of the Holy Quran, Almadinah Almunawwarah, Saudi Arabia.
5. ElHadj, Y.O.M., I.A. AlSughayeir, A.M. Khorsi and 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, pp: 72-84.
6. Buckwalter, T., 2004. Buckwalter Arabic Morphological Analyzer. Version 2.0. http://students.cs.byu.edu/~jonsafar/buckwalter.html
7. Dror, J., D. Shaharabani, R. Talmon and S. Wintner, 2004. Morphological analysis of the Quran. Literary Linguist. Comput., 19: 431-452. DOI: 10.1093/llc/19.4.431
8. Talmon, R. and S. Wintner, 2003. Morphological tagging of the Quran. Proceedings of the Workshop on Finite-State Methods in Natural Language Processing, Apr. 2003, Budapest, Hungary, pp: 1-8. http://cs.haifa.ac.il/~shuly/publications/talmon-wintner-eacl03.pdf
9. ElHadj, Y.O.M., I.A. Al-Sughayeir and A.M. Al-Aansari, 2009. Arabic part-of-speech tagging using the sentence structure. Proceeding of the 2nd International Conference on Arabic Language Resources and Tools, Apr. 2009, Cairo, Egypt, pp: 241-245. http://www.elda.org/medar-conference/pdf/5.pdf
10. 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
11. Freeman, A., 2001. Brill’s POS tagger and a morphology parser for Arabic. Proceeding of the ACL’01 Workshop on Arabic Language Processing.
12. Maamouri, M. and C. Cieri, 2002. Resources for Arabic natural language processing at the LDC. Proceeding of the International Symposium on the Processing of Arabic, Tunisia, 2002, pp: 125-146.
13. Diab, M., K. Hacioglu and D. Jurafsky, 2004. Automatic tagging of Arabic text: From raw text to base phrase Chunks. Proceeding of the HLTNAACL’04, pp: 149-152.
14. Banko, M. and R.C. Moore, 2004. Part of speech tagging in context. Proceeding of the 20th international conference on Computational Linguistics, Aug. 23-27, Association for Computational Linguistics Morristown, New Jersey, USA., Article No. 556. http://portal.acm.org/citation.cfm?id=1220435
15. Tlili-Guiassa, Y., 2006. Hybrid method for tagging Arabic text. J. Comput. Sci., 2: 245-248. http://www.scipub.org/fulltext/jcs/jcs23245-248.pdf
16. Al Shamsi, F. and A. Guessoum, 2006. A hidden markov model-Based POS Tagger for Arabic. http://www.cavi.univ-paris3.fr/lexicometrica/jadt/jadt2006/PDF/004.pdf
17. Brants, T., 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. http://portal.acm.org/citation.cfm?id=974178