Journal of Computer Science 10 (9): 1865-1873, 2014
ISSN: 1549-3636

doi:10.3844/jcssp.2014.1865.1873 Published Online 10 (9) 2014 (http://www.thescipub.com/jcs.toc)

ARABIC PART OF SPEECH TAGGING USING K-NEAREST NEIGHBOUR AND NAIVE BAYES CLASSIFIERS COMBINATION

Rund Mahafdah, Nazlia Omar and Omaia Al-Omari

Center For Artificial Intelligence Technology, Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia, 43600, Bangi, Selangor, Malaysia

Received 2014-03-02; Revised 2014-03-13; Accepted 2014-05-02

ABSTRACT

Part Of Speech (POS) tagging forms the important preprocessing step in many of the natural language processing applications such as text summarization, question answering and information retrieval system. It is the process of classifying every word in a given context to its appropriate part of speech. Different POS tagging techniques in the literature have been developed and experimented. Currently, it is well known that some POS tagging models are not performing well on the Quranic Arabic due to the complexity of the Quranic Arabic text. This complexity presents several challenges for POS tagging such as high ambiguity, data sparseness and large existence of unknown words. With this in mind, the main problem here is to find out how existing and efficient methods perform in Arabic and how can Quranic corpus be utilized to produce an efficient framework for Arabic POS tagging. We propose a classifiers combination experimental framework for Arabic POS tagger, by selecting two best diverse probabilistic classifiers used in numerous works in non-Arabic language; namely K-Nearest Neighbour (KNN) and Naive Bayes (NB). The Majority voting is used here as the combination strategy to exploit classifiers advantages. In addition, an in-depth study has been conducted on a large list of features for exploiting effective features and investigating their role in enhancing the performance of POS taggers for the Quranic Arabic. Hence, this study aims to efficiently integrate different feature sets and tagging algorithms to synthesize more accurate POS tagging procedure. The data used in this study is the Arabic Quranic Corpus, an annotated linguistic resource consisting of 77,430 words with Arabic grammar, syntax and morphology for each word in the Holy Quran. The highest accuracy in the results achieved is 98.32%, which can be a significant enhancement for the state-of-the-art for Arabic Quranic text. The most effective features that yield this accuracy are a combination of w_0 (the current word), p_0 (POS of the current word), p_{-3} (POS of three words before), p_{-2} (POS of two words before) and p_{-1} (POS of the word before).

Keywords: Part of Speech, Natural Language Processing, Classification

1. INTRODUCTION

Part Of Speech (POS) disambiguation is the ability to computationally figuring out which POS of a word is activated by its use in a certain context. Additionally, it can be explained as the procedure of determining a suitable POS tag for every single word in a sentence. Fine-grained POS (morpho-syntactic or morphological) tagging is the procedure of determining POS, tense, number, gender and other morphological information for every single word in a sentence (Feldman, 2006; Schmid and Laws, 2008). POS tagging is an essential language analysis task in almost all NLP systems, including information extraction, corpus annotation

Corresponding author: Rund Mahafdah, Center For Artificial Intelligence Technology, Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia, 43600, Bangi, Selangor, Malaysia

projects, word-sense disambiguation and etc. The next step is another high-level language analysis task by which the output of POS taggers will be generally submitted to. Both syntactic parsing (Mohamed, 2010) and Named Entity Recognition (Benajiba, 2009) are included in these high-level language analyses.

Part of speech tagging is a crucial NLP problem. It entails a large amount of challenging problems including different kinds of unknown words and POS ambiguities.

Such words that could not be found neither in the dictionary nor in the training corpus are known as "Unknown Words". To understand the meaning of a sentence of unknown words is more essential than known words. They also carry more semantic information than known words (Vadas and Curran, 2005). Unknown words are part of the open POS classes like verbs and nouns and it is not probable to be in the closed classes like particles. In fact, the sources of open-ended text, including web corpus provide NLP systems with major challenge unknown words (Weischedel *et al*., 1993).

Natural languages are naturally ambiguous (Tomita, 1985; Dukes *et al*., 2010). Ambiguity is most likely to occur at various levels of the Natural Language Processing (NLP) task (Dandapat, 2009; Jurafsky *et al*., 2009). In the case where the ambiguity shows up in one word is referred to as lexical ambiguity like POS ambiguity (Manning and Schutze, 1999).

2. RELATED WORK

POS tagging provides essential information about word forms used in sentences of natural language. Utilizing this information varies depending on the specific NLP application (i.e., information retrieval, machine translation), in which it is used.

As depicted in **Fig. 1**, there are two techniques in POS tagging; linguistic taggers and machine learning approaches. Machine learning approaches are divided into two main groups; supervised and unsupervised.

2.1. Linguistic Taggers

Linguistic-based taggers specify the relevant knowledge as a set of rules or constraints that is done by linguists. These models generally require years of work as they are ranging from a few hundred to several thousand rules. Research in automated POS tagging began in the midst 60 and 70's (Klein and Simmons, 1963; Harris, 1962; Greene and Rubin, 1971). Researchers manually established rules for tagging.

2.2. Machine Learning Approaches

The POS disambiguation may be seen as a classification problem: The tag set is the classes and an automatic classification method used in each repetition of a word to one class based on the evidence from the context. Picking up the classification method is the most critical phase in POS disambiguation. Machine learning field is the origin of the majority of the recent approaches (Navigli, 2009). The methods of machine learning vary from methods with fully unsupervised to fully supervised methods.

However, unsupervised and supervised approaches differ greatly. Some of the most important differences are shown in **Table 1**.

Arabic is a Semitic language which is spoken by more than 450 million people. It is also an extremely derivational and firmly structured language. Moreover, Arabic is among the six official languages of the United Nations. It is grammatically ambiguous.

Unfortunately, there have been no open sources available POS tagger that are designed especially for Arabic to handle the community's dependence on fundamental NLP tools. Besides, due to the difficulty with the Arabic POS disambiguation problems and the limitations of the existing work in the literature, thus, the Arabic POS disambiguation problems need more investigations. To date, little research has been done in the area of statistical NLP for Arabic, which is confined by having less openly accessible manually annotated corpora. To be able to minimize the huge cost of manually developing annotated corpora, the progress of the POS taggers is of substantial value.

2.3 Arabic Part of Speech

According to Haywood and Nahmad (1962), Arabic words can be classified into three main POS. Later, these POS will be again categorized into more detailed POS. The three main parts of speech are:

2.3.1. Noun

A noun in Arabic is a name or a word that describes a person, thing, or idea. Usually the noun group is divided into sub-group of derivatives (i.e., nouns derived from verbs, nouns derived from other nouns and nouns derived from particles) and primitives (nouns not so derived). These nouns could be further sub-categorized by number, gender and case. This category contains what would be categorized as participles, pronouns, relatives, demonstratives and interrogatives.

Table 1. The main differences between unsupervised and supervised methods

Unsupervised	Supervised
Induction of the tag set	Selection of the tag set
Use untagged training data	Use pre-tagged training data
Induction of the training data	Creation of dictionaries using a tagged corpus
Domain independent: It has the ability to speedily scale to any language	Domain dependent: Its performance can drop substantially when test data comes from a different domain
It theoretically has worse performance	It may be more accurate especially

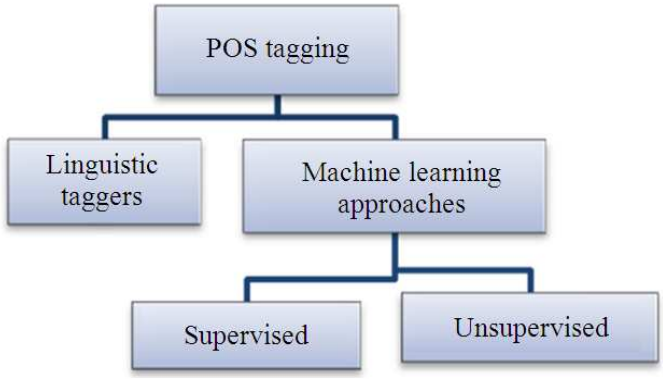

Fig. 1. Classification of POS tagging models

2.3.2. Verb

The verb classification in Arabic is similar to English, although the tenses and aspects are different. The verbs can be sub-categorized by 'type' (perfect, imperfect, imperative), person, number and gender and the tag name reflects this sub-category. As an example, the word كسرتم ksrtm "you [plural, masculine] broke" is a perfect verb in the second person masculine plural form. An indicative imperfect second person feminine singular verb such as تكتبين tktbyn "you [singular, feminine] are writing".

2.3.3. Particle

The particle group contains: Prepositions, adverbs, conjunctions, interrogative particles, exceptions (these are consisting of the Arabic words that are equivalent to the word except and the prefixes non-, un- and im-.) and interjections.

The group of particle contains adverbs, conjunctions and prepositions. All of these can be found in Arabic either as individual words or as clitics that come with the next word. Other particles are interjections, exceptions and negative particles.

2.4. POS Tagging Approaches used for Arabic

The amount of study of POS tagging has been done on Arabic language with different dialects. Each of these dialects has its own small number of vocabularies. Mohamed (2010) described that "Arabic POS tagging is still in the stage of research since Arabic poses different problems than those posed by English."

The problems of Arabic studies in POS tagging are as follows (El-Hadj, 2009; Al Gahtani *et al*., 2009):

- It experiences the knowledge acquisition bottleneck problem
- Arabic is a language with a complicated morphology which raises the number of unknown words
- The problem of lack of resources which are even rarely or not freely open for research, for instance lexicons
- Arabic dialects are seldom written which makes annotated corpora and lexicons to be hardly developed
- Regarding to some reasons, including the lack of writing short vowels, Arabic is among the languages with a high degree of ambiguity

Based on the literature of Arabic POS tagging, there are many approaches have been proposed for such aim. These approaches are based on different assumptions and rules and have had different accuracy results in contributing to the field. Some of the most related research on the POS tagging approaches which have been done for Arabic are summarized in the **Table 2**.

Table 2. A summary of POS tagging approaches for Arabic

Approach	Author	Accuracy %
Transformation-based	Freeman and McVea (2001)	---
Transformation-based + morphological analyzer	Al Gahtani *et al*. (2009)	96.10
SVM	Diab and Habash (2007) Diab *et al*. (2004)	95.49
SVM + morphological analyzer	Habash and Rambow (2005)	97.60
Statistical	Mohamed (2010)	94.37
Statistical + rule based	Khoja (2001)	90.00
Memory based learning	Yang *et al*. (2007)	91.50
Rule based + memory based	Tlili-Guiassa (2006)	85.00
HMM	AL-Shamsi and Guessoum (2006)	97.00
HMM with morphological Analyzer	El-Hadi *et al*. (2009)	96.00
HMM with morphological Analyzer with lexicon	Mansour *et al*. (2007)	96.12
Classifier + regular expressions	Kulick (2010)	95.15
MAXPOST+ TBL+ TnT	Albared et al. (2009)	96.50

3. ARABIC POS TAGGING FRAMEWORK

In this section, we propose a solution for Arabic POS tagging framework which is the classifiers combination of the best supervised machine learning-based taggers including K-Nearest Neighbour (KNN) and Naïve Bayes (NB). They are combined using majority voting algorithms. Classifiers combination with machine learning individuals is effectively used on several languages and typically outperform their individuals. As well as the need of an Arabic analysis tool (Diab and Habash, 2007), we are attempting to discover how the mentioned techniques can be used in Arabic and what are they gained results. We are going to combine the best of the classifiers in order to earn benefit of every single method.

3.1. Corpus and Pre-processing

In this study we have used the Quranic Arabic corpus in our approach. The Quranic corpus is preprocessed prior to the experiments, starting with tokenization. Tokenization can be defined as the process of splitting out words (morphemes) from running text (Jurafsky *et al*., 2009). It is an essential and an initial step in NLP. Splitting sentences into tokens is the purpose of tokenization. It also enables them to end up being given into POS tagger or a morphological analyzer for further processing (Attia, 2007).

Quran is the Islamic religious book and it is written in classical Quranic Arabic (in 600 CE). According to Dukes and Habash (2010) and Dukes *et al*. (2010), the Quranic Arabic corpus is an annotated linguistic resource that indicates the Arabic syntax, grammar and morphology for every single word in the Quran. The research project is structured at the University of Leeds by computing research group within the School of Computing (http://corpus.quran.com). Arabic Quranic Corpus is composed of 77,430 words. The corpus is a reference with numerous levels of analysis consisting of POS tagging, morphological segmentation. Every single word of the Quran is tagged using its POS along with several morphological features.

In this phase, the researchers acquire Quranic Arabic verses for preliminary tokenization. After that, the automatically tokenized text will be examined manually and then corrected. Manual correction includes manual normalization of the tokenized text.

3.2. Features Selection

Here are three different kinds of feature from the sliding window:

3.2.1. Word Features

It includes word form n-grams, typically unigrams, bigrams and trigrams suffice. As well as, the sentence last word that refers to a punctuation mark (’.’, ’?’, ’!’) is important. Different word features used in this experiment.

3.2.2. POS Features

Annotated Parts Of Speech (POS) and ambiguity classes n-grams. Regarding words, considering unigrams, bigrams and trigrams is enough. The ambiguity class for a specific word ascertains when POS is possible.

3.2.3. Affix and Orthographic Features

They consist of prefixes and suffixes, capitalization, hyphenization and similar information related to a word form. They are simply employed to signify unknown words. **Table 3** indicates a rich feature set of the experiment.

3.3. The Combined Classifiers

The following phase of the workflow is the combined classifiers. Two classifiers have been used in the combination, namely K-Nearest Neighbour (KNN) and Naive Bayes (NB). On one hand, the K-Nearest Neighbour algorithm will assist when the test pair has similar characteristics to one of the training examples. On the other hand, NB is selected because it is known to obtain high performance.

3.3.1. K-Nearest Neighbour classifier

The K-Nearest Neighbour (KNN) is a well-known instance-based classifier. KNN is referred as a powerful method to the various text classification problems (Duda *et al*., 2001; Yang, 1994). Additionally, KNN is known as lazy learners, because it defers the decision on how to generalize beyond the training data until every new query instance is experienced. In the KNN algorithm, a new input instance needs to be part of the same class as its K nearest neighbours in the training dataset. After that when a new input instance is classified in the class of K nearest neighbours between all training instances. The "closeness" is identified as a distance metric, such as the Euclidean distance.

3.3.2. Naive Bayes

The Naive Bayes (NB) classifier is a well-known machine learning technique. It is an uncomplicated probabilistic classifier determined by utilizing Bayes' theorem (from Bayesian statistics) having strong (naive) independence assumptions. The detailed word for the fundamental probability model could be an independent feature model. Simply a Naive Bayes classifier presumes that the presence (or absence) of a specific feature of a class (that is attribute) is unrelated to the presence (or absence) of any other feature.

3.4. Voting Algorithms (Combination Strategies)

The selection algorithm as the center of this methodology ascertains the accuracy of the combined classifiers. It does it by finding the right answer provided a set of three answers. A number of the selection algorithms includes: Majority (simple voting), plural (total) voting, tag precision, stacking (cascade classifiers).

Majority voting is the most straightforward voting technique. It looks at only the most probable class given by every single classifier then it finds the most repeated class label among this crisp output set. Weighted majority voting as a trainable variant of majority voting which increases every single vote by a weight before the actual voting. The weight for every classifier could be gained; for instance by calculating the classifiers' accuracies on a validation set. Another voting technique is board count which considers the whole n-best list of a classifier, not only the crisp 1-best candidate class.

3.5. Evaluation

In general, the evaluation measures in classification problems are defined from a matrix with the numbers of examples correctly and incorrectly classified for each class, named confusion matrix. The confusion matrix for a binary classification problem (which has only two classes, positive and negative), is shown in **Table 4**.

The FP, FN, TP and TN concepts may be described as:

- False Positives (FP): Instances predicted as positive, which are from the negative class
- False Negatives (FN): Instances predicted as negative, whose true class is positive
- True Positives (TP): Instances correctly predicted as pertaining to the positive class
- True Negatives (TN): Instances correctly predicted as belonging to the negative class

Table 3. Rich feature pattern set used in the experiment and its symbol

Word features	w_{-3}, w_{-2}, w_{-1}, w_0, w_{+1}, w_{+2}, w_{+3}
POS features	p_{-3}, p_{-2}, p_{-1}, p_0, p_{+1}, p_{+2}, p_{+3}
Prefixes	s_1, s_1s_2, $s_1s_2s_3$, $s_1s_2s_3s_4$
Suffixes	s_n, s_{n-1} s_n, s_{n-2} s_{n-1} s_n, s_{n-3} s_{n-2} s_{n-1} s_n
Binary word features	All upper case, all lower case, contains a number
Word length	Integer

Table 4. Confusion matrix

	Predicted class	
True class	Positive	Negative
Positive	TP	FN
Negative	FP	TN

The evaluation measure most used in practice is the accuracy rate (ACC). By its percentage of correct predictions, it evaluates the effectiveness of the classifier. Accuracy equation is computed as follows:

$$ACC = ((TP+TN)/ (TP+TN + FP+FN)) *100$$

4. RESULTS AND EVALUATION

This section demonstrates the results of the experiments performed on the Quranic Arabic corpus by applying the identified individual classifiers as well as selected combinations. A sample of experimental results will be delicately elaborated. Furthermore, the classifiers and a list of features that lead to the best result will be stated out.

4.1. Experiment Test Set

The Quranic Arabic corpus includes syntactic and morphological annotation of the Quran and builds on the verified Arabic text distributed by the Tanzil project (Tanzil.net). It consists of 77,430 words. The researchers of the present study performed their experiment based on the whole Quran corpus. For each experiment, the whole words and a random set of features of those words are chosen from the corpus.

4.2. Experimental Results

The experiment applied the 28 features as it is explained in **Table 5**, including the word and its part of speech, word features (7), POS features (7), prefixes (4), suffixes (4), binary word features (3) and word length (1) on the datasets of the Arabic Quran. For each individual experimental run, a random set of features was chosen as well as a single classifier or a combination. The total conducted runs are 138 within the experiment. The percentage of the total score for each classifier and the supplemented set of features are calculated and the highest accuracy obtained is 98.32%. The best classifier that gives such accuracy is a combination of NB and KNN. The set of features is a combination of w_0 (the current word), p_0 (POS of the current word), p_{-3} (POS of three words before), p_{-2} (POS of two words before), p_{-1} (POS of the word before) and p_0 (POS of the current word).

4.2.1. Individual Classifiers Approach

4.2.1.1. Naive Bayes (NB)

Table 6 shows the list of the highest results obtained by applying the NB classifier along with different sets of feature patterns. The table shows the highest accuracy obtained which is 91.77%, by set 14.

4.2.1.2. K-Nearest Neighbour (KNN)

Table 7 shows the list of the highest results achieved by applying the KNN classifier combined with different sets of features patterns. The table shows the highest accuracy obtained which is 95.5%, by set 14.

4.2.2. Combined Classifiers Approach

Table 8 illustrates a list of the highest results obtained by applying the combination of KNN and NB classifiers as well as different sets of features patterns. The table shows the highest accuracy obtained which is 98.32%, by set 14.

Table 5. Feature pattern set used in the experiments

Feature symbol	Feature pattern
F1	w_0
F2	p_0
F10	p_{-3}
F11	p_{-2}
F12	p_{-1}
F13	p_0
F14	p_{+1}
F15	p_{+2}
F16	p_{+3}
F17	s_1
F18	$s_1 s_2$
F19	$s_1 s_2 s_3$
F20	$s_1 s_2 s_3 s_4$
F21	s_n
F22	$s_{n-1} s_n$
F23	$s_{n-2} s_{n-1} s_n$
F24	$s_{n-3} s_{n-2} s_{n-1} s_n$
F25	Contains a number
F26	All upper case
F27	All lower case
F28	Integer

Table 6. The highest accuracy percentages % achieved by NB

Set no.	Features	ACC
Set3	F1, F2, F21, F22, F23, F24, F25, F26 and F27	89.75
Set13	F1, F2, F17, F18, F19, F20, F21, F22, F23, F24, F25, F26, F27 and F28	89.89
Set14	F1, F2, F10, F11, F12 and F13	91.77
Set18	F1, F2, F21, F22, F23 and F24	89.90

Table 7. The highest accuracy percentages % achieved by KNN

Set no.	Features	ACC
Set2	F1, F2, F17, F18, F19, F20, F21, F22, F23 and F24	80.38
Set13	F1, F2, F17, F18, F19, F20, F21, F22, F23, F24, F25, F26, F27 and F28	89.24
Set14	F1, F2, F10, F11, F12 and F13	95.50
Set16	F1, F2, F17, F18, F19, F20, F25, F26, F27 and F28	87.62

Table 8. The highest accuracy percentages % achieved by the combination of KNN and NB

Set no.	Features	ACC
Set13	F1, F2, F17, F18, F19, F20, F21, F22, F23, F24, F25, F26, F27 and F28	90.89
Set14	F1, F2, F10, F11, F12 and F13	98.32
Set16	F1, F2, F17, F18, F19, F20, F25, F26, F27 and F28	87.75
Set21	F1, F2, F25, F26 and F27	94.25

Finally, the result of the study revealed that the proposed model is a significant enhancement for the state-of-the-art for Arabic POS tagging. The research results were compared with the latest researches on Arabic POS tagging and have proved higher accuracy.

By taking advantage of combining classifiers and by evaluating the set of results obtained each time by applying a classifier with a set of features, the highest accuracy was 98.32% achieved by KNN and NB combination. Besides, the most effective feature that accomplish this accuracy is a combination of namely; w_0 (the current word), p_0 (POS of the current word), p_{-3} (POS of three words before), p_{-2} (POS of two words before), p_{-1} (POS of the word before) and p_0 (POS of the current word).

5. CONCLUSION

Arabic is considered as a widely spoken language that is being spoken by approximately 450 million people, what makes it as the fourth widespread language. However, in the computer world and especially the Internet content, Arabic language only represents 3.00% of the overall Internet's lingual content. Moreover, using Arabic in computerized systems is an issue nowadays because of the complex morphology and structure of such a language.

As has been said before, this research mainly contributes to the field of POS tagging and is specified for the Arabic language. The set of contributions can be achieved, in particular and in general by the research are as follows:

- The research has studied, examined and presented a set of rich feature patterns that assist in enhancing the POS tagging especially in rich morphological languages such as Arabic
- The research has presented a model that significantly enhances the performance of POS tagging in Arabic based on the combination of classifiers and integration a set of rich feature patterns
- The model contributes in improving the disambiguation of the word category and grammatical tagging in Arabic language

As a future work, we believe that improving the features and patterns for tags is a possible strategy to raise the accuracy levels of POS tagging systems. They also intend to perform further investigation on this POS tagging approach in order to reduce the error rate and apply it as a basis for a parsing and analyzing system framework.

6. REFERENCES

Al Gahtani, S. W. Black and J. McNaught, 2009. Arabic part-of-speech-tagging using transformation-based learning. The University of Manchester.

Albared, M., N. Omar, and M.J.A. Aziz, 2009. Arabic part of speech disambiguation: A survey. Int. Rev. Comput. Soft., 4: 517-532.

Al-Shamsi, F. and A. Guessoum, 2006. A hidden markov model-based POS tagger for Arabic. University of Sharjah.

Attia, M., 2007. Arabic tokenization system. Proceedings of the Workshop on Computational Approaches to Semitic Languages: Common Issues and Resources, (CIR' 07), ACM, USA, pp: 65-72.

Benajiba, Y., 2009. Arabic named entity recognition. PhD Thesis, Universidad Politécnica de Valencia Valencia, Spain.

Dandapat, S., 2009. Part-of-speech tagging for Bengali. M.Sc. Thesis, Indian Institute of Technology, Kharagpur, India.

Diab, A.A., B. Teulat-Merah, D. This, N.Z. Ozturk, and D. Benscher et al., 2004. Identification of drought-inducible genes and differentially expressed sequence tags in barley. Thior. Applied Genet., 109: 1417-1425. PMID: 15517148

Diab, M. and N. Habash, 2007. Arabic dialect processing Tutorial. Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics. (ACL' 07), Association for Computational Linguistics Stroudsburg, PA, USA., pp: 5-6.

Duda, R.O., P.E. Hart and D.G. Stork, 2000. Pattern Classification. John Wiley and Sons, New York, USA, ISBN-10: 0471056693, pp: 680.

Dukes, K. and N. Habash, 2010. Morphological annotation of quranic Arabic. University of Leeds.

Dukes, K., E. Atwell and M. Sharaf, 2010. Syntactic annotation guidelines for the quranic Arabic dependency Treebank. Proceedings of the 7th International Conference on Language Resources and Evaluation, May 19-21, European Language Resources Association, pp: 1-6.

El-Hadi, T., A. Oujilal, M. Boulaich, L. Sqalli, and M. Kzadri, 2009. Plemorphic adenoma of the infratemporal space: A new case report. Int. J. Otolaryngol., 2009: 529350-529352. DOI: 10.1155/2009/529350

El-Hadj, Y.O.M., 2009. Statistical part-of-speech tagger for traditional Arabic texts. J. Comput. Sci., 5: 794-800. DOI: 10.3844/jcssp.2009.794.800

Feldman, A., 2006. Portable language technology: A resource-light approach to morpho-syntactic tagging. PhD Thesis, The Ohio State University, USA.

Freeman, R.E. and J. McVea, 2001. A stakeholder approach to strategic management. University of Virginia.

Greene, B.B. and G.M. Rubin, 1971. Automatic grammatical tagging of English. 1st Edn., Brown University, Providence, Rhode Island, pp: 306.

Habash, N. and O. Rambow, 2005. Arabic tokenization, part-of-speech tagging and morphological disambiguation in one fell swoop. Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, (ACL' 05), Association for Computational Linguistics Stroudsburg, PA, USA., pp: 573-580. DOI: 10.3115/1219840.1219911

Harris, Z., 1962. String analysis of language structure. Int. J. Am. Linguist., 30: 415-420.

Haywood, J.A. and H.M. Nahmad, 1962. A New Arabic Grammar of the Written Language. 2nd Edn., Lund Humphries Publishers Ltd, London, pp:687.

Jurafsky, D., J.H. Martin and A. Kehler, 2009. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition. 2nd Edn., Prentice Hall, New Jersey, ISBN: 9780130950697.

Khoja, S., 2001. Thematic indexing in video databases. PhD Thesis, University of Southampton.

Klein, S. and R. Simpson, 1963. A computational approach to grammatical coding of English words. J. ACM, 10: 334-347. DOI: 10.1145/321172.321180

Kulick, S., 2010. Simultaneous tokenization and part-of-speech tagging for Arabic without a morphological analyzer. Proceedings of the ACL Conference Short Papers, (CSP' 10), Association for Computational Linguistics Stroudsburg, PA, USA., pp: 342-347.

Manning, C.D. and H. Schutze, 1999. Foundations of Statistical Natural Language Processing. MIT Press Cambridge ed. MA, USA. ISBN-10: 0262133601.

Mansour, J.K., T. Mateus and R.C.L. Lindsay, 2007. Disguise effects on identification accuracy from sequential and simultaneous lineups.

Mohamed, E., 2010. Orthographic enrichment for Arabic grammatical analysis. PhD Thesis, Indiana University, Indiana, United States.

Navigli, R., 2009. Word sense disambiguation: A survey. ACM Comput. Surveys, 41: 1-69. DOI: 10.1145/1459352.1459355

Schmid, H. and F. Laws, 2008. Estimation of conditional probabilities with decision trees and an application to fine-grained POS tagging. Proceedings of the 22nd International Conference on Computational Linguistics, (CCL' 08), Association for Computational Linguistics Stroudsburg, PA, USA., pp: 777-784.

Tlili-Guiassa, Y., 2006. Hybrid method for tagging arabic text. J. Comput. Sci., 2: 245-248. DOI: 10.3844/jcssp.2006.245.248

Tomita, M., 1985. An efficient context-free parsing algorithm for natural languages. Proceedings of the 9th International Joint Conference on Artificial Intelligence, (CAI' 85), Morgan Kaufmann Publishers Inc. San Francisco, CA, USA., pp: 756-764.

Vadas, D. and J. Curran, 2005. Tagging unknown words with raw text features. Proceedings of the Australasian Language Technology Workshop, (LTW' 05), Sydney, Australia, pp: 32-39.

Weischedel, R., R. Schwartz, J. Palmucci, M. Meteer and L. Ramshaw, 1993. Coping with ambiguity and unknown words through probabilistic models. Computational Linguistics - Special issue on using large corpora: II. 19: 361-382.

Yang, X., H.J. Mo, F.C. Van Den Bosch, A. Pasquali and C. Li *et al*., 2007. Galaxy groups in the SDSS DR4. I. The catalog and basic properties. Astrophys. J., 671: 153-170. DOI: 10.1086/522027

Yang, Z., 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate. J. Mol. Evol., 39: 306-14. PMID: 7932792