Part Of Speech Tagging For Arabic Text Based Radial Basis Function

Osama R.Shahin [1†] and Rady El Rwelli[2],

[1]Department of Computer Science, Jouf University, Gurayat, Saudi Arabia
[*]Physics and Mathematics Department, Faculty of Engineering, Helwan University, Egypt
[2]Department of Arabic Languages, Jouf University, Gurayat, Saudi Arabia

Summary
The progress of peoples and their advancement linked to the extent of their keenness to spread their civilization and scientific progress in their language and tongue. Language is the pot of civilization without a doubt. The Arabic language is one of the most spoken languages in the world, with more than 467 million people speaking it. The Arabic language has succeeded in playing a unique civilized role, making it the leader of civilization and knowledge at the level of science for centuries in a row. In the current era, and because of the proliferation of modern technologies, advanced technology and software that take the foreign languages its own platform, the role of the Arabic language has declined in influence, and it has become imprisoned in universities, schools and the role of science that only its people can learn. However, at the present, it is time to gather its forces to meet the requirements of the present and the future in the technical, civilizational and epistemological field. The aim of this work is to detect the Part of Speech (POS), in which to simplify the rules of the Arabic language to ensure its vitality by keeping pace with the language of technical developments. In addition, to spread Arabic language sciences through the application resulting from this reality via the Internet, which makes it the greatest influence in simplifying its own rules for non-native speakers and the desire to study and learn its arts. The proposed model consists of three phases, that starting by the phase of reading the text and segment it into sentences and words, followed by the phase of analyzing and tagging each word inside the text. Finally, the syntax analyzer (Parser) that used to determine the (POS) will built by using the Radial Basis Function network (RBF) which is a type of Artificial Neural Network (ANN). RBF simply contains three layers (layer that responsible for entering data, the hidden layer in which processing done, and adjusting the network weights to perform their role in matching the inputs and outputs, and the output layer).

Keywords:
Natural Language Processing, Part of Speech, Syntax analyzer, Radial Basis Function network

1. Introduction

The impact of technological developments has extended to all aspects of life, including the Arabic language. These modern developments and technologies have affected the Arabic language with a positive and negative effect. One of the positive effects of the technology is the emergence of a set of modern electronic tools and applications that are concerned with the correct rules of spelling and sound pronunciation. Among the negative effects especially after the spread of social media the emergence of a mixture of languages written in English and read in Arabic, which called (Franco-Arab) leads to the departure of words from what was set for it. For this and other negatives, the Arab person, especially researchers, must take care of it and harness all modern technology means to preserve it, develop information systems to deal with the Arabic text, and maintain its existence from moral and semantic fissures.

The Arabic language is one of the Semitic languages, which considered one of the most spoken languages in the world. Where more than 467 million people speak it [1-3], in addition to that 1.5 billion Muslims around the world need to use it in their various affairs, regarding reading the Holy Quran and understanding its meanings and other Worship and rituals. The Arabic language derives its permanence and spread from being the language of the Noble Qur'an. Allah Almighty has chosen for his book the most eloquent languages. Therefore, the association of the Arabic language with the Qur'an made it to remain alive and preserved by its preservation [4,5]. The Arabic language demonstrated in more than one field its enormous ability to interact with the influences of all times and even influence them. Perhaps the decision of the United Nations General Assembly on December 18, 1973 AD to include the Arabic language in the official languages of the General Assembly and its main committees is evidence of the impact of this language on civilizations and peoples.

Artificial intelligence is the behavior and specific features of computer programs that make them match the human mental capabilities and working methods [6-8]. Among the most important of these features are the ability to learn, reasoning, logical thinking and understanding natural languages. Artificial intelligence is concerned with creating algorithms that simulate human intelligence. One of the important areas of artificial intelligence is making the machine learning capable of learning on its own, which makes it able to classify the analysis of feelings through some of the words indicating this in addition to categorizing opinions on a specific topic across social media platforms [9]. One of the most important branches of machine learning is Natural Language Processing (NLP), which aims to understand and treat natural languages to extract science

Manuscript received November 5, 2025
Manuscript revised November 20, 2025
https://doi.org/10.22937/IJCSNS.2025.25.11.9

and knowledge from them. Among the most important levels of natural language analysis (morphological, grammatical, semantic analysis) [10].

In this work, the proposed model consists of three phases, that starting by the phase of reading the text and segment it into sentences and words, followed by the phase of analyzing and tagging each word inside the text. Finally, the syntax analyzer (Parser) that used to determine the (POS) will built by using the Radial Basis Function network (RBF) [11,12]. After experiments, this system demonstrated proficiency in extracting POS. The confusion matrix used here to demonstrate the efficiency and accuracy of the proposed algorithm.

2. Related Work

Most of the recent research were focused on the vocabulary of the Arabic language is concentrated in four areas. The first field is in identifying the parts of speech (POS), secondly, is in the analysis of sentiment analysis and opinions, thirdly some applications used to replace Arabic words with similar meaning images to indicate better meanings, especially for children, and finally, plagiarism detection in Arabic documents have been developed consider the fourth field . Here we review the latest research in these fields.

Here, the work was concern on extracting parts of speech (POS) in the Arabic, Berber, and Romanian languages was reviewed by using a vector machine (SVM). Results of each of the 25 proposals being reviewed, along with a review of the error ratios and the accuracy of each algorithm; In order to motivate and guide researchers to do more research on applications interested in the Arabic language due to the scarcity of the work targeting them [1]. In this work, the researchers have proposed an effective approach to extract the parts of the sentence written in Arabic using the bee colony algorithm (ABC). Where the words of the sentence are represented as drawn paths. Then the algorithm begins to scan the path of each word while recording the length of this path, and comparing it with the lengths of the word paths in the database of 18 million words; this is to determine what this word is [13]. In this work, the effect of initial processing of Arabic words on the performance of text classification was studied using machine learning algorithms - deep learning (DL). Including removing the hamzah on the thousand and two points on the bound T character, which reduces the search time in the database and then reflected in the speed of the performance of the algorithm [14].

Due to the importance of the work in the field of distinguishing Arabic words, the work in this paper focused on identifying Arabic words through the comments of YouTube users.

Moreover, know their opinions and their participation on the subject of the video watched. A set of machine learning algorithms (ML) were developed for data / text mining, natural language processing and linguistic classification to discover the Arabic comparative text [15]. In this work, recent studies were reviewed and the methodologies of studies for distinguishing and analyzing the words of the Arabic text were compiled using machine learning applications with a greater focus on the Holy Qur'an which will help those interested in the topic of extracting and understanding Arabic texts in pursuing their future studies [16].

On the other hand, the researchers here have proposed two systems to help specialists in detecting plagiarism in natural Arabic texts. The first method relied on word weights for measuring semantic similarity relationships between text words. The second approach relies on machine learning mechanisms (ML), where lexical, grammatical and semantic features have been combined to assist in the task of detection using automated learning techniques such as support vector machines (SVM), decision trees (DT) and random forests (RF) [17]. In this work, the study dealt with a set of research papers that applied to extract parts of the sentence in the Arabic language using the Markov model (MC).

The focus here was on the mechanisms for converting words written in Arabic into words written in English letters but with the same Arabic pronunciation (transliteration); for ease of working with English characters in most programming languages.

In addition, to avoid diacritical marks that can make classification processes difficult, precision ratios compared between several algorithms to be a starting point for researchers to improve the performance of these algorithms and for the accuracy of the desired results [18].

In this paper, the researchers have provided an algorithm to convert Arab children's stories into representative pictures that can effectively explain the meaning of words. Such works are a major challenge to machine learning algorithms. The proposal has been applied in three stages. The first of these stages is the application of natural language processing techniques to analyze the text in the story. To extract keywords for all characters and events in each sentence. Second, the algorithm implements an image captioning process through a pre-trained Deep Learning Model (DL) for all database images. Finally, using similarities between the words of the phrases extracted in the first stage and the titles of the pictures in the database in the second stage, most of the pictures in the sentence words are retrieved [19].

In this work, a set of features has been chosen that can represent and distinguish a group of Arabic words, starting with the features based on the spelling of the word and the features based on the word roots. This work used 4900 texts were used to train and test the proposed algorithm in addition to using machine-learning techniques (SVM) as well as the inverse document repeat mechanism (LTC). In addition, to using different metrics to measure the performance of the algorithm [20]. Finally, attention has been paid to studying the extraction of the sentence written in the Gulf dialect, as it includes work on preparing a set of data and signs to discover and classify words in addition to applying two methods of machine learning, namely, the vector machine (SVM) and short-range bi-directional memory (Bi-LSTM)[21] .

3. Proposed Method

Treating Arabic vocabulary and sentences is a difficult task when compared to treatments based on other languages. This is due to the most Arabic words are contains connected letters and contain many suffixes and precedents. Here in this work we present a complete system needed to extract parts of speech i.e. verb, letter and noun. As mentioned before, the proposed algorithm consists of two phases. These phases are text-preprocessing phase, tagging words phase, and classification model.

3.1 Text Preprocessing

In the first phase, which is the phase of entering the text written in the Arabic language, basic processing operations for the text begin, such as determining the beginnings and ends of all the text sentences. This is done by assuming that the sentences in the Arabic language end with a point (.). Followed by the process of word separation (all pronouns from the word) e.g. separation the letter "هـــ" from the word "كتبه", so it will write as "هـــ" + "كتب". Also, separate punctuation marks such as commas, question marks, exclamation marks, etc. by removing them from the words; The Lexical Analyzer makes it easy in the second stage to tag and classify each word. For the purpose of Arabic, word segmentation, here A CRF model will be used [22].

3.2 Words Tagging

Here, the process of tagging the words is done by placing two letters to indicate whether the word is a name, verb or letter, for example the noun - the noun with its derivatives or any letter related to the noun ... etc. will be suggested to tag the it (NN).

The verb will be suggested to tag with (PZ) it if the verb is a past. In addition, the proposed tag will be (FZ) if the verb is a present. However, in the case of the imperative verb the proposed tag is (QA). The tagging for a prepositions will be (HC), and for the letters for "جزم" the proposed tag will be (HM), while it will be suggested that the affix be marked as (HP), and for the interrogative letters (HQ). However, the verb sentence will tagged by "VP", and the noun sentence will tagged by "NP".

3.3 Classification Model and Lexicon

This phase is concerned with searching for a word in the Lexicon. Here the Lexicon will be built at this stage by means of Radial Basis Function network (RBF), and then the proposed neural network will be trained on a group of words, the number of these words reached 20 thousand words taken from the dataset called (Kalimat) [23]. The Lexicon here will assumed as adynamic Lexicon due to its ability to accommodate a new set of words that will be supplied in training dataset, i.e. add a new corpus. The syntax analyzer (Parser) that used to determine the (POS) will built by using the Radial Basis Function network (RBF) which is a type of Artificial Neural Network (ANN). RBF simply contains three layers (layer that responsible for entering data, the hidden layer in which processing done, and adjusting the network weights to perform their role in matching the inputs and outputs, and the output layer).

The result of this training is a model that is able to distinguish words alone. Figure 1 shows a diagram of the proposed algorithm.

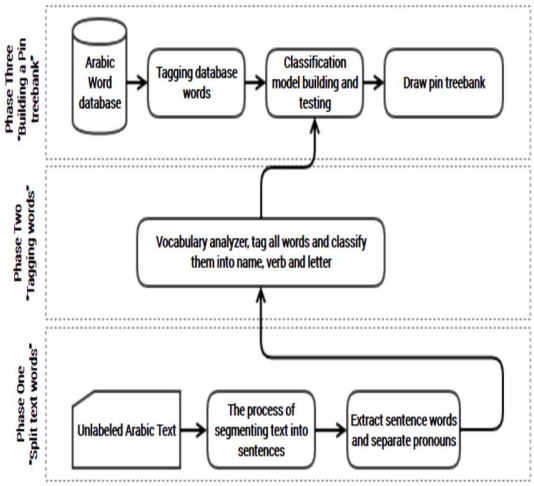

Fig. 1. Phases of the proposed algorithm

3. Results and Discussions

As mentioned previously, the corpora collected from KALIMAT [12], which consider the famous Arabic natural language resource. Table 1 summarized the contents of this corpus.

Table 1: Description of KALIMAT corpus.

Type of the Document	Number
Arabic articles	20,291
Document Summaries.	22,348
Recognized articles.	20,291
Speech Tagged articles.	20,291
Morphologically articles.	20,291

Here we 18,000 words (Noun, verb, and letter), this corpora are categorized into two groups; the first group used for training, and the second group used for testing. In the beginning, all words from each group arranged in text files with UTF-8 encoding. The words collected statistics summarized in Table 2.

Table 2: Corpus statistics

Word	The total number
Noun	8000
Verb	4000
Letter	6000

Figure 2 gives an example, which shows how an Arabic sentence is tagged by using the proposed algorithm. The input is the sentence "يلعب أحمد بالكرة". it contains four words, the first word is a present verb which tagged by "FZ". The next word is a noun which tagged by "NN". The last word is preposition which tagged by "HC". Finally the noun "الكرة" that tagged by "NN". The root of the sentence tagged by "VP", due to the type of the sentence, which is a verb sentence.

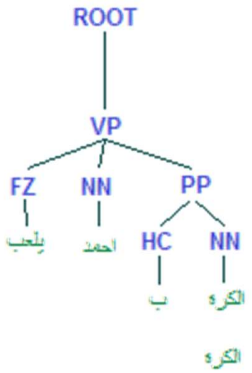

Fig. 2. Sentence tagging using proposed algorithm

The performance of the proposed algorithm will obtained by aiding of the confusion matrix. Here, the algorithm's accuracy in recognizing the noun, verb and letter will be measured separately. Table 3 shows the different performance factors that used to build in such matrix.

Table 3: Performance measures

Confusion Matrix	Test Result (Positive or Negative)	
Actual speech unit (True)	True Positive (TP)	False Positive (FP)
Actual speech unit (False)	False Negative (FN)	True Negative (TN)

After the essential performance factors explained measured, the following metrics will be used to measure the classification performance measures: Accuracy, precision, recall, and F1 score. The accuracy metric can be calculated by using eq. 1 as follows:

$$Accuracy=(TN+TP)/(TN+TP+FP+FN) \quad (1)$$

While, the Precision can be determined by eq. 2:

$$Precision=TP/(TP+FP) \quad (2)$$

The formula of Recall metric given by eq. 3:

$$Recall=TP/(TP+FN) \quad (3)$$

Finally, the F1 score is measured weighted by recall and precision, and it is determined by eq. 4:

$$F1\text{-}Score=(2\times Preision\times Recall)/(Preision+Recall) \quad (4)$$

Table 4 shows the results of the algorithm in identifying the (POS) in terms of classification performance measures.

Table 4: Results of the proposed algorithm

POS	Accuracy	Precision	Recall	F1-Score
Noun	92.45%	93.22%	92.44%	92.83%
Verb	92.94%	93.27%	93.87%	93.57%
Letter	90.67%	92.01%	91.18%	91.59%

The performance measures of the proposed algorithm in comparison with the other implemented algorithms are depicted in Table 5.

Table 5. Comparison of different reliable methodologies of pos tagging of arabic text

Author (s), Ref. No.	Model / Methodology	Corpus	Accuracy
Kadim, A. and Lazrek [24]	Viterbi algorithm	Nemlar Arabic	75.38%
KhetamYassen, MajdiSawalha, and Fawaz Al Zaghoul [25]	Bigram tagger	BAQ	90.79%
Ba-Alwi, Fadl Mutaher, Mohammed Albared, and Tareq Al-Moslmi [26]	HMM with prefix guessing	Holy Quran	88.1%
Ibrahim, Hossam S., Sherif M. Abdou, and Mervat Gheith [27]	MSA / BL	Modern Standard and Egyptian dialectal	95%
Badaro, Gilbert, Ramy Baly, Hazem Hajj, Nizar Habash, and Wassim El-Hajj [28]	Sentiment lexicons and RBF	College-educated native speakers of Arabic	80%

5. Conclusion

This article presents a method for identifying the parts of an Arabic sentence using using RPF. The algorithm consists of three basic phases, the first phase starting by entering the text written in the Arabic language. Basic processes will occurred on this text such as identifying the beginnings and ends of all the sentences of the text. The second phase will begins to recognize and tag the words of each sentence and remove all the extra spaces through the text vocabularies. Finally, a grammatical parser will built to define the sentence architecture. The results of this work indicate that the RPF model yielded satisfactory results when used to design the parser to show the sentence structure.

References

[1] Yousif, J., & Al-Risi, M. (2019). Part of Speech Tagger for Arabic Text Based Support Vector Machines: A Review. ICTACT Journal on Soft Computing.

[2] Habib, Sandy. "The Exponents of Eleven Simple, Universal Concepts in Three Semitic Languages." International Journal of Arabic Linguistics 6, no. 1-2 (2020): 68-90.

[3] Ameur, Mohamed Seghir Hadj, Farid Meziane, and Ahmed Guessoum. "Arabic Machine Translation: A survey of the latest trends and challenges." Computer Science Review 38 (2020): 100305.

[4] Saad, Hasbollah Bin Mat. The Principles Of Shari'ah: Texts And Materials. Pena Hijrah Resources, 2020.

[5] Wright, Thomas. The Life of Sir Richard Burton. BoD–Books on Demand, 2020.

[6] Wan, Shaohua, Zonghua Gu, and Qiang Ni. "Cognitive computing and wireless communications on the edge for healthcare service robots." Computer Communications 149 (2020): 99-106.

[7] Alomari, M., and M. Jabr. "The effect of the use of an educational software based on the strategy of artificial intelligence on students' achievement and their attitudes towards it." Management Science Letters 10, no. 13 (2020): 2951-2960.

[8] Meshrif Alruily and Osama R. Shahin "Sentiment Analysis of Twitter Data for Saudi Universities." nternational Journal of Machine Learning and Computing (IJMLC) 10 (2020): 18-24.

[9] Winkler, David A. "Role of Artificial Intelligence and Machine Learning in Nanosafety." Small 16, no. 36 (2020): 2001883.
[10] Li, Yan, Manoj A. Thomas, and Dapeng Liu. "From semantics to pragmatics: where IS can lead in Natural Language Processing (NLP) research." European Journal of Information Systems (2020): 1-22.
[11] Zhao, Zhitao, Yang Lou, Yifeng Chen, Hongjun Lin, Renjie Li, and Genying Yu. "Prediction of interfacial interactions related with membrane fouling in a membrane bioreactor based on radial basis function artificial neural network (ANN)." Bioresource technology 282 (2019): 262-268.
[12] Kansa, E. J., and P. Holoborodko. "Fully and sparsely supported radial basis functions." International Journal of Computational Methods and Experimental Measurements 8, no. 3 (2020): 208-219.
[13] Alhasan, A., & Al-Taani, A. T. (2018). POS tagging for arabic text using bee colony algorithm. Procedia computer science, 142, 158-165.
[14] Elnagar, A., Einea, O., & Al-Debsi, R. (2019). Automatic text tagging of Arabic news articles using ensemble deep learning models. In Proceedings of the 3rd International Conference on Natural Language and Speech Processing (pp. 59-66).
[15] Alharbi, F. R., & Khan, M. B. (2019). Identifying comparative opinions in Arabic text in social media using machine learning techniques. SN Applied Sciences, 1(3), 213.
[16] Salloum, S. A., AlHamad, A. Q., Al-Emran, M., & Shaalan, K. (2018). A survey of Arabic text mining. In Intelligent Natural Language Processing: Trends and Applications (pp. 417-431). Springer, Cham.
[17] Cherroun, H., & Alshehri, A. (2018, April). Disguised plagiarism detection in Arabic text documents. In 2018 2nd International Conference on Natural Language and Speech Processing (ICNLSP) (pp. 1-6). IEEE.
[18] Yousif, J. (2019). Hidden Markov Model Tagger for Applications Based Arabic Text: A review. Journal of Computation and Applied Sciences IJOCAAS, 7(1).
[19] Zakraoui, J., Elloumi, S., Alja'am, J. M., & Yahia, S. B. (2019). "Improving Arabic text to image mapping using a robust machine learning technique." IEEE Access, 7, 18772-18782.
[20] Al-Thubaity, A., Alqarni, A., & Alnafessah, A. (2018, April). "Do Words with Certain Part of Speech Tags Improve the Performance of Arabic Text Classification?." In Proceedings of the 2nd International Conference on Information System and Data Mining (pp. 155-161).
[21] Alharbi, R., Magdy, W., Darwish, K., Abdelali, A., & Mubarak, H. "Part-of-speech tagging for Arabic Gulf dialect using Bi-LSTM. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation." (LREC 2018).
[22] Lafferty, John, Andrew McCallum, and Fernando CN Pereira. "Conditional random fields: Probabilistic models for segmenting and labeling sequence data." (2001).
[23] https://sourceforge.net/projects/kalimat/
[24] Kadim, Ayoub, and Azzeddine Lazrek. "Parallel HMM-based approach for arabic part of speech tagging." Int. Arab J. Inf. Technol. 15, no. 2 (2018): 341-351.
[25] KhetamYassen, MajdiSawalha, and Fawaz Al Zaghoul. "Part-Of-Speech Tagging For Classical And Msa Arabic Text Using Nltk." New Trends in Information Technology (NTIT)–2017 (2017): 106.
[26] Ba-Alwi, Fadl Mutaher, Mohammed Albared, and Tareq Al-Moslmi. "Choosing the Optimal Segmentation Level for POS Tagging of the Quranic Arabic." Current Journal of Applied Science and Technology (2017): 1-10.
[27] Ibrahim, Hossam S., Sherif M. Abdou, and Mervat Gheith. "Sentiment analysis for modern standard arabic and colloquial." arXiv preprint arXiv:1505.03105 (2015).
[28] Badaro, Gilbert, Ramy Baly, Hazem Hajj, Nizar Habash, and Wassim El-Hajj. "A large scale Arabic sentiment lexicon for Arabic opinion mining." In Proceedings of the EMNLP 2014 workshop on arabic natural language processing (ANLP), pp. 165-173. 2014.