Rule Based Approach for Arabic Part of Speech Tagging and Name Entity Recognition

Mohammad Hjouj Btoush
Department of Computer Science
Al-Balqa' Applied University
Al Salt, Jordan

Abdulsalam Alarabeyyat
Department of Software Engineering
Al-Balqa' Applied University
Al Salt, Jordan

Isa Olab
Department of Computer Science
Al-Balqa' Applied University
Al Salt, Jordan

***Abstract*—The aim of this study is to build a tool for Part of Speech (POS) tagging and Name Entity Recognition for Arabic Language, the approach used to build this tool is a rule base technique. The POS Tagger contains two phases:The first phase is to pass word into a lexicon phase, the second level is the morphological phase, and the tagset are (Noun, Verb and Determine). The Named-Entity detector will apply rules on the text and give the correct Labels for each word, the labels are Person(PERS), Location (LOC) and Organization (ORG).**

Keywords—POS; Speech tagging; Speech recognition; Text phrase; Phrase; NLP

I. INTRODUCTION

Natural Language Processing (NLP) is a field of computer science, artificial intelligence, and linguistics concerned with the interactions between computers and human (natural) languages. NLP is related to the area of human–computer interaction. Many challenges in NLP involve natural language understanding; that is, enabling computers to derive meaning from human or natural language input, while others involve natural language generation [1].

The objective of this study is to contributetothe existing literature body of building atoolfor POS and Name Entity Recognition for Arabic language.

1) Part of speech

Part of speech tagging (POST) is also known as POS tagging, word classes, morphological classes,lexical tags or just tagging as a process that aims to assign a specific tag to each word of a sentence to indicate the function of that word in a specific context ,the suitable tag is chosen from a set of tags based on some rules, examples of part of speech are: nouns ,verbs, pronouns, prepositions ,adverbs and other tags, the results of part of speech are for many applications such as speech recognition, natural language parsing,information retrieval, information extraction, question answering, text to-speech conversion, machine translation, grammar correction and many more [2].

POS tagging is one of the tools and the components for infrastructure to Natural Language Processing of a given language. POS tagging is necessary in many fields such as: text phrase, syntax, semantic analysis and translation [3].

2) POS-tagging techniques

There are many techniques that may be used separately or with each other for tagging words to its classes ,the most famous methods are Rule-based, stochastic and transformation (hybrid) method:

Rule-Based Tagging: Algorithm for assigning part of speech based on two stages, the first stage uses a lexicon (dictionary) to assign each word a list of potential parts of speech, the second stage uses a huge list of hand-written disambiguation rules to winnow down this list to a single part of speech for each word [2].

Stochastic Tagging: The idea behind this approach is to pick the most likely tag for this word, in this method we need a tagged corpus to get probabilities about the ambiguous word, hidden Markov model algorithm is an example of the this technique [2].

Transformation-Based Tagging: The is an approach which combines Rule-based and Stochastic technique ,like the Rule-based, TB is based on rules to specify the tag for each ambiguous word but like Stochastic Tagger, in which rules are automatically induced for data [2].

3) Name Entity Recognition

Named Entity Recognition (NER) is a sub problem of information extraction and involves processing structured and unstructured documents and identifying expressions that refer to peoples, places, organizations and companies. NER is a fundamental task and it is the core of natural language processing (NLP) system. NER involves two tasks, which are: firstly the identification of proper names in text, and secondly the classification of these names into a set of predefined categories of interest, such as: persons,organizations, locations, date and time expressions [4].

a) Name Entity Recognition techniques

Hand-made Rule-based: focuses on extracting names using lots of human-made rules set. Generally the systems consist of a set of patterns using grammatical syntactic and Orthographic features in combination with dictionaries [4].

Machine Learning-based: NER system, the purpose of Named Entity Recognition approach is converting identification problem into a classification problem and employs a classification statistical model to solve it. In this type of approach, the systems look for patterns and relationships into text to make a model using statistical models and machine learning algorithms. The systems identify and classify nouns into particular classes using machine learning algorithms [4].

Hybrid Based: this technique is combined the rule based and machine learning-based and make new methods using strongest points from each method[4].

II. Literature Review

Mohamed and Kubler (2010) presented a method for POS tagging for Arabic language without segmentation which would be unrealistic for naturally occurring Arabic. Developed approach which used the full POS tagset. the resultsfor experiments suggest that the segmentation isn't very important in POS tagging, reached the best results in tagging for known and unknown word, the worst result of the segmentation-based approach is its low accuracy on unknown words [5].

Hadni et al (2013) proposed an efficient and accurate POS Tagging technique for Arabic language by using hybrid approach. Due to the ambiguity issue, Arabic Rule-Based method suffers from misclassified and unanalyzed words. To overcome these two problems, they presented a Hidden Markov Model (HMM) matched with Arabic Rule-Based method. The proposed technique used different contextual information for the words, this method tested with two corpora: the Holy Quran Corpus and Kalimat Corpus for undiacritized Classical Arabic language [6].

Elhadj(2009) 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. His project related to the computerization of the Holy Quran which was to build a textual corpus of the Holy Quran, he 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. Each tag in this system is used to give all possible state of the HMM and the transitions between tags 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 their developed tag set [7].

Umansky (2010) presented a web-based algorithm for the task of POS tagging of unknown words (words appearing only a small number of times in the training data of a supervised POS tagger). If a sentence is containing an unknown word that tagged by a trained POS tagger, this algorithm collects from the web contexts that are partially similar to the context of an unknown word in sentence, which are then used to compute new tag assignment for unknown word, this algorithm enables fast multi-domain unknown word tagging [8].

Aboaoga and Ab Aziz (2013) presented a rule-based approach for recognition Arabic Named Entity. The goal of this paper is to use the rule based approach for recognizing the named entities that include person names in economic, politic and sport domain. The method consists of three main steps: pre-processing, automatic named entity tagged and applying the rules. The method had been applied on Arabic corpus of three domains (politic, economic and sport) to recognize the named entity (person name) in the text. Then, the evaluation method has been used to compute the performance measure for each domain [9].

Oudah and Shaalan (2013) presented a hybrid approach for name entity recognition, and this technique contains two levels: The first level is the rule base which used to produce name entity labels based on lists of name entity keywords and contextual rules.The second level is the Machine Learning based intended to make use of rule-based component's name entity decisions as features aiming at enhancing the overall performance of the name entity recognition task [10].

Elsebai et,al (2009) presented a rule based approach for name entity recognition, the system consists of two components the first is GATE (General Architecture for Text Engineering) and the second is BAMA (Buckwalter Arabic Morphological Analyzer) the GATE used to perform tokenization and then annotate the text by highlighting those words that belong to the Introductory Verb List and Introductory word List lists [11].

Then the BAMA used to perform the input word and returns the stem rather than the root then they used a set of keywords to guide them to the phrases that probably include person names [11].

III. System Architecture

1) POS Tagger

In this project we have used the rule base to tagging the Arabic text, the proposed system(as shown below) consists of two phases : The first phase is the lexicon analyzer which contains all Arabic particles including prepositions, adverbs, conjunctions, interrogative Particles, exceptions, and interjections. The second phase is a morphological phase which uses morphological information such as the patterns of the word and its affixes (such as prefixes, suffixes and infixes) to presume the class of the words.

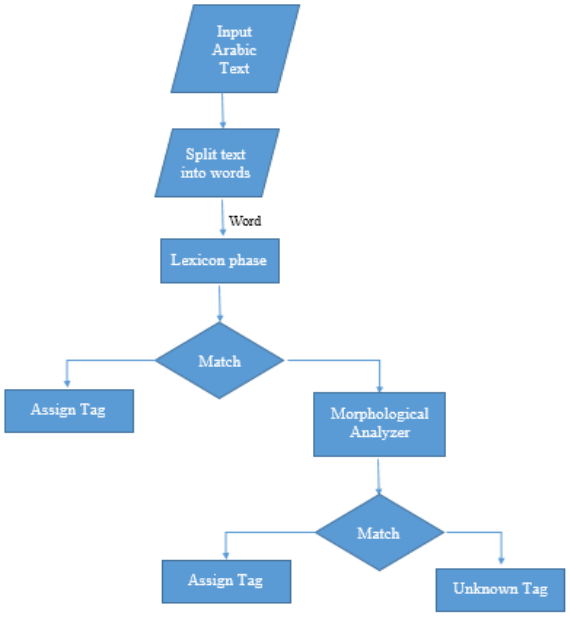

Fig. 1. The Architecture of the Tagging System

As shown in figure 1 the tagging system reads an Arabic text then splits it into words after that it takes every word and put it into the first level (Lexicon phase), in this case if it exists we return the corresponding tag, if not; the word is transferred to the second phase (Morphological phase). When completing processing the word, if it either matches, it returns the presumed tag; otherwise the tag is Unknown.

a) Lexicon phase

In this phase and after the text is split into words,the lexicon holds all Arabic fixed words and particles (prepositions, adverbs, conjunctions, interrogative particles, exceptions, questions and interjections), and every word is searched in the lexicon if it is found, the corresponding tag is returned, if it is not found in the lexicon; then it moves to the next phase (Morphological phase).

b) 1.2 The Morphological Phase

As we know, the Arabic language is more complicated because it has the largest number of possible affixes (especially prefixes and suffixes), Arabic possesses, and a large number of derivational.

In the Arabic language there are many signs that indicate if the word is a noun or a verb, some patterns in words are used with verbs and others are used with nouns, and some patterns are used for both verbs and nouns.

The information from affixesdoesn not help us to determine the exact word classification within the two major categories nouns and verbs. Certain prefixes, suffixes or infixes come with certain classes of words, the followings are some of hand-written rules to classify a word into a noun or a verb or Particles:

Rule 1: the following prefixes "كال", "بال", "فال", "وال" if it comes in the beginning of a word map it refers to Noun class.

Rule 2: the following suffixes "ائي","ائك","ائه","اؤك","اؤه","اءك","اءه","هما","كما" if it comes in the end of a word map itrefers to Noun Class.

Rule 3: the following prefixes "سي","ست","سن","سأ","سا","لا","لأ","لن","لت","لي" if it comes in the end of a word map itrefers toVerb class.

Rule 4: the following suffixes "ي", "ه", "ك", "ا", "ن", "و" if it comes in the end of a word map itrefers to Verb class.

Rule 5: if the word has the pattern (فعلى, فعول, فعاء) map it to Noun class.

Rule 6: if the word ends with "ات" map it to Noun class.

Rule 7: if the word end with "ون, ين"but starts with "ي" or "ن" map the word to Verb class.

Rule 8: if the word ends with "ون, ين" and doesn't start with "ي" or "ن" map the word to Noun class.

Rule 9: if the word has the pattern(مفاعل,مفعيل,مفعال,مفتعل,منفعل,مفعول,متفعل,مفعلل) map it to Noun class.

Rule 10: if the word has the pattern (فعاليل, أفاعيل, مفاعيل, فواعيل) map it to Noun class.

Rule 11: if the word has the pattern (استفعل, افعوعل) map it to Verb class.

2) The Name Entity Detector

The second subsystem in our project is the Name Entity Recognition, we have used the rule base approach to label the Arabic text, each word is labeled with any of the three labels Person (PERS), Location (LOC) and Organization (ORG).

The data set is three lists, the first is for person names and contain 4030 word, the second is for location names and contain 2193 words, the third is for organization names and contain 268 words, thearchitecture of the name entity system is shown in the figure below:

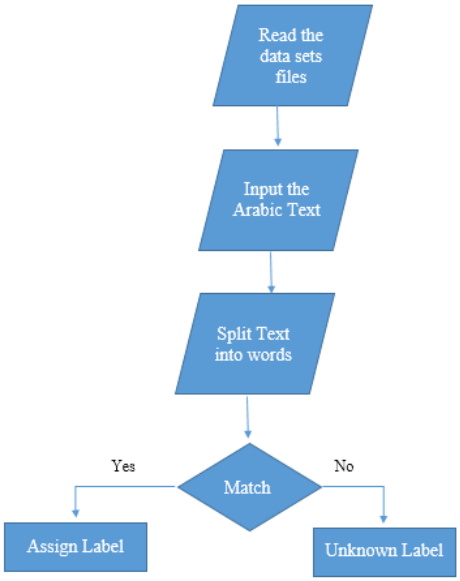

Fig. 2. The Architecture of the Name Entity System.

First the system reads the data sets that contain the person, location, organization names and stores these names, then the system reads the Arabic text and splits it into words

After that the system will apply the rules on the words and if they match then the label will be assigned to that word, if not, the label will be Unknown.

Here are some examples how the name entity detector works:

If we take the country name "المملكة الاردنية الهاشمية" this country name consists of three words so the system takes this name as one word give it LOC label.

Another example is the "الوكالة الدولية للطاقة الذرية" this organization name consists of four words, the system will take this name as one word and give it ORG Label.

If we take the person name "فلاديمير بوتين" this person name consists of two words the system take this name as one words and give it PERS Label.

If we take the person name "خالد بن الوليد"this person name consists of two name and separate them with the word "بن" so

the system checks if two person name separate between them with " بن" and give it the PERS Label.

IV. Experimental Results

First the POS tagger was tested on a file that contains 793 word; the result shows that the tagger successfully tagged 679 words. However, the name entity detector was tested on a file that contains 490 words and successfully tagged 480 word.

Here som examples of the POS Tagger and name entity detector :

TABLE I. Some of Result by POS Tagger

الطيبون يغرسون الجمال داخل الروح لا اراديا	الطيبون_N يغرسون_V الجمال_N داخل_Nالروح_N لا_DETاراديا_N
ليس الفخر أن تقهر قويا بل الفخر أن تنصف ضعيفا	ليس_Vالفخر_N أن_DETتقهر_V قويا_N بل_DETالفخر_N أن_DETتنصف_V ضعيفا_N
عبادالرحمنالذينيمشونعلىالأرض	عباد_N الرحمن_N الذين_Nيمشون_N على_DETالأرض_N
لنتنالواالبرحتىتنفقوامماتحبون	لن_DETتنالوا_V البر_N حتى_DETتنفقوا_V مما_N تحبون_V
علمتنيالدنياأنأعفوعنالمخطئ	علمتني_V الدنيا_N أن_DETأعفو_V عن_DETالمخطئ_N

TABLE II. Some of Result by Name Entity Detector

صائب عريقات يقول ان الاصرار الفلسطيني على التصويت في مجلس الامن كان رسالة الى واشنطن	صائب عريقات-PERS يقول ان الاصرار الفلسطيني على التصويت في مجلس الامن-ORG كان رسالة الى واشنطن-LOC
صحيفة الغارديان البريطانية ترى ان ازمات 2015 السياسية والامنية والاقتصادية ستكون مترابطة ولا يمكن فصلها في عصر العولمة والتقلبات	صحيفة الغارديان-ORG البريطانية ترى ان ازمات 2015 السياسية والامنية والاقتصادية ستكون مترابطة ولا يمكن فصلها في عصر العولمة والتقلبات
تعرصت الاثار السورية في محافظة ادلب شمال سوريا للدمار والتخريب	تعرصت الاثار السورية في محافظة ادلب-LOC شمال سوريا-LOC للدمار والتخريب
السفير الفلسطيني في الامم المتحدة يؤكد تسليم بلاده وثيقة للانضمام للمحكمة الجنائية الدولية ويقول ان فلسطين ستلجا للخيار القانوني لمحاسبة اسرائيل	السفير الفلسطيني في الامم المتحدة-ORG يؤكد تسليم بلاده وثيقة للانضمام للمحكمة الجنائية الدولية ويقول ان فلسطين-LOCستلجا للخيار القانوني لمحاسبة اسرائيل-LOC
أفاد مراسل الجزيرة في لبنان بمقتل ثلاثة من عناصر حزب الله وعدد من جنود الجيش السوري في مواجهات مع جبهة النصرة في محيط بلدة فليطة السورية المحاذية للأراضي اللبنانية.	أفاد مراسل الجزيرة-ORG في لبنان-LOC بمقتل ثلاثة من عناصر حزب الله-ORG وعدد من جنود الجيش السوري في مواجهات مع جبهة النصرة في محيط بلدة فليطة السورية المحاذية للأراضي اللبنانية

V. Conclusions and Future Work

In this paper the Rule Based Approach used for the Part Of Speech Tagging and Name Entity Recognition were tested. The POS Tagger contains two phases first the lexicon phase and the second Morphological phase, the name entity applies rules on Arabic text to extract person names, location and organization and give for each of them their labels. Future work will focus in increasing the tag set of the tagger and increasing the labels of the name entity detector, as well as the rules of tagger and Name Entity detector.

VI. References

[1] B Bataineh and E.Bataineh, An Efficient Recursive Transition Network Parser for Arabic Language, *Proceedings of the World Congress on Engineering* 2009 Vol II WCE 2009, July 1 - 3,(2009), London, U.K.

[2] J.H Martin and D.Jurafsky, D., Speech and language processing., *International Edition*(2010).

[3] A. AL-Taani and S.Abu Al-Rub, A Rule-Based Approach for Tagging Non-Vocalized Arabic Words, *The International Arab Journal of Information Technology*, Vol. 6, Issue. 3(2009).

[4] A.Mansouri, S.Lilly Affendey and A.Mamat, Named Entity Recognition Approaches, *International Journal of Computer Science and Network Security*, VOL.8 Issue.2,(2008).

[5] E.Mohamed and S.K¨ubler , "Is Arabic part of speech tagging feasible without word segmentation?." Human Language Technologies: *The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, (2010).

[6] M.Hadni, S. Ouatik, A.Lachkar and M.Meknassi, Hybrid Part-Of-Speech Tagger for Non-Vocalized Arabic Text, *International Journal on Natural Language Computing (IJNLC)* Vol. 2, Issue.6, (2013).

[7] Y.Elhadj.Statistical Part-of-Speech Tagger for Traditional Arabic Texts,*Journal of Computer Science Vol.5, Issue* 11, pp.794-800, (2009) ,ISSN 1549-3636,© 2009 Science Publications.

[8] S.Umansky-Pesin, R.Reichart and A.Rappoport,, A multi-domain web-based algorithm for POS tagging of unknown words. *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*. Association for Computational Linguistics, (2010).

[9] M.Aboaoga and M. Ab Aziz, "Arabic Person Names Recognition By Using A Rule Based Approch", *Journal of Computer Science*, Vol. 9, Issue 7, pp. 922-927, (2013) ISSN: 1549-3636© 2013 Science Publications.

[10] M.Oudeh and K.Shaalan, K., Person Name Recognition Using Hybrid Approach, *lecture Notes n Computer Science*, *Natural Language Processing and Information Systems*, *Springer Berlin Heidelberg*, vol. 7934, pp. 237-248.

[11] A.Elsebai, F. Meziane and F.Belkredim, A Rule Based Persons Names Arabic Extraction System, *Communications of the IBIMA*Vol.11, (2009) ISSN: 1943-7765.