ISSN 1996-3343

Asian Journal of Applied Sciences

http://knowledgiascientific.com

Knowledgia SCIENTIFIC
A Place to Publish Outstanding Research

Asian Journal of Applied Sciences 7 (1): 13-21, 2014
ISSN 1996-3343 / DOI: 10.3923/ajaps.2014.13.21

Named Entity Recognition for Political Domain in Arabic Language

[1]Halema H. Mhamed Alshref and [2]Mohd. Juzaiddin Ab Aziz
[1]Department of Computer Science, Faculty of Al-tahadi Sirte, Al-Jufrah University, Waddan, Libya
[2]School of Computer Science, Faculty of Information Technology, University Kebangsaan Malaysia, 43600 Bangi, Selangor, Malaysia

Corresponding Author: Mohd. Juzaiddin Ab Aziz, School of Computer Science, Faculty of Information Technology, University Kebangsaan Malaysia, 43600 Bangi, Selangor, Malaysia

ABSTRACT

Named Entity Recognition (NER) began in late 1991 with a small number of general categories such as names of persons, names of organizations and names of locations. This work describes the development and implementation of Arabic Named Entity Recognition System (ANER system) for the Arabic Language. For identification and classification Named Entities (NEs) in the text such as persons, locations, organizations and temporal values. NER plays a significant role in various types of Natural Language Processing (NLP) applications, especially in information extraction, information retrieval, machine translation, syntactic parsing/chunking and question-answering. The NER task was considerably more challenging when it was targeting a morphologically complex language such as Arabic due to its complexity. The Arabic language has some peculiarities which harden the NER task. Arabic has a rich and complex morphology. The main aim of this research was to use the rule based approach to design and implement an Arabic NER system for the political domain. The rule based approach consisted of a lexicon, in the form of verb contextual clue lists and the noun contextual clue list, together with a set of grammar rules which were responsible for recognizing and classifying NEs. Considering the system evaluations tested on ANER corpus. Taken human annotated corpus and evaluated the system and then compared the results against this corpus. The accuracy was 94.86% of ANER system, the results showed that the approach achieved an accuracy of 82.76% for Person NE, 98.3% for Location, 100% for Organization and 98.37% for MISC.

Key words: Information extraction, Arabic named entity recognition, named entities, natural language processing, Arabic language, political domain

INTRODUCTION

NER began in the late 1991's with the small number of general categories such as names of persons and organizations and locations (Sekine, 2004; Nadeau and Sekine, 2007). Named entity recognition is one of the multiple tasks in extracting important information which involves identification and classification of words. This includes the classification of persons (names, adjectives), organization (companies, government organizations, etc.), locations (countries, cities, rivers, etc.) and temporal values (date, time and numbers). NE refers to a proper name in the Arabic language. It indicates a name of a specific and self-contained entity. It also means that it is a noun representing unique entities and which are not shared by others, for example: احمد (Ahmed), السعودية (Saudi Arabia), ليبيا (Libya), (Company) (Benajiba *et al*., 2008). A NE in the

Arabic language exists in many types. Recognize that developing a system to extract important data from documents becomes essential. However, the Arabic language has its own characteristics and dealing with the Arabic language is a complicated task. Determining the entities named in the Arabic language is very difficult because they do not begin with capital letters, thus, making it difficult if not impossible to establish their distinction in a text based only on the first letter of words (Abuleil, 2004). It is important to develop an algorithm that would be able ensure the recognition of NEs in situations of exponential complexity for clarity. Also, there are many names in the Arabic language where there is no limit on what adjectives or nouns are to be used as NE. The difficulty lies in how to solve the huge ambiguity between different ANE types and how to formalize distinctions between names, adjectives and verbs as well as how to identify proper nouns which are persons, locations, organizations or temporal values (Kaiser and Miksch, 2005). The proposed method of this study is to use the rule-based approach. The rule-based approach is the first form of technology used by researchers in NER. This approach is based on the written rules of human experts. It is one of the methods that were used to extract the NER of the text field (Moisiadis *et al*., 2008). In the early days, systems were based primarily on the NER pattern-matching rules as based on the use of dictionaries and rules established by hand and the use of a set of keywords as guides to the phrases that probably include NEs. We have used the approach, rule-based approach is the first technology used by researchers in NER. This approach is based on the written rules of human experts. It is one of the methods that were used to extract the NER of the text field (Moisiadis *et al*., 2008). In the early days, systems were based primarily on the NER pattern-matching rules as based on the use of dictionaries and rules established by the hand and we used a set of keywords to guide us to the phrases that probably include NEs. Our approach utilizes contextual and POS tagging information to classify NEs. The context is represented by means of surrounding words (contextual clue words) which are used as clues for each NE type (Attia and Rashwan, 2004). The POS tagging information is represented by the POS of the surrounding words, the pos of the NEs themselves and also the POS patterns of the NEs that do not include proper nouns.

In case of Arabic language, there has been some work on ANER. Shaalan and Raza (2009) developed the system, person NER for Arabic, using a rule-based approach. The system consists of a dictionary, in the form of gazetteer name lists and a grammar, in the form of regular expressions which are responsible for recognizing named entities. The components of the name in the Arabic language and divided them into different constitutional elements. Arabic has well-defined naming practices.

Zaghouani *et al*. (2010) developed the system, Arabic information extraction system that is used to analyze large volumes of news texts every day to extract the NE type's person, organization, location, date and number. The approach used to solve problems in Arabic language substantially differs from those used for Indo European and Finno-Ugric languages (Vergyri *et al*., 2004). The solution proposed to overcome these constraints was creating various lexicons of person names, organizations, location etc. following the rule-based approach. Use the lexicon entries to match directly the NEs. Also use the lexicon entries in their local grammar to match partially unknown entities. Used local grammar to detect potential NEs not included in their lexicon and also built a list of Trigger words and Stop words. This approach is motivated by the characteristics of Arabic language. It is more practical and easy to implement when compared to machine learning approaches that usually require large tagged corpora and dictionaries. Europe Media Monitor (EMM) system is optimized for rule-based systems (Pouliquen *et al*., 2005). The main aim of this research is to propose a Rule-based Approach for the Arabic NE in text and classified as different

types of NEs. In order to achieve the main objective the research also, develop rules to identify the main features of NEs in Arabic political documents. Design and implement an Arabic NER using the Rule-based approach. Compare the performance of our system with those of human systems.

Components proper noun: Targeting proper nouns represent one of the important tasks in information extraction which involves the identification and classification of words or sequences of words denoting a concept or entity. As Fig. 1 summarizes the types of proper noun.

Structure of the NE: A NE in the Arabic language exists in many types. Table 1 illustrates some of the proper names and each one will be accompanied with an example.

Proper nouns in Arabic are usually either nouns or adjectives used as proper nouns and there are no formal distinctions. Many of the most common Arabic names, e.g., "احمد" or Ahmed (Eng. most praise-worthy), "محمد" or Mohammed (Eng. more praise-worthy) and "كريم" or Cream (Eng. immortal) are just adjectives and are still used as such. Personal names can even be noun phrases such as "نور الهدي" or Nur Alhuda, (Eng. the light of guidance) and "نور الدين" or Nur Aldin (Eng. the soundness of religion). However "Information" can exist as components following a noun or adjective and the sentences in Table 2 illustrate the types of Noun. Table 2 illustrates a list of types of the information for the most commonly used Arabic and English sentences.

This is some example show the type of proper noun in Table 3. This type may be a person name as example 1 in Table 4, location name as 2 in Table 4, organization name as 3 in Table 4, or temporal value as 4 in Table 4.

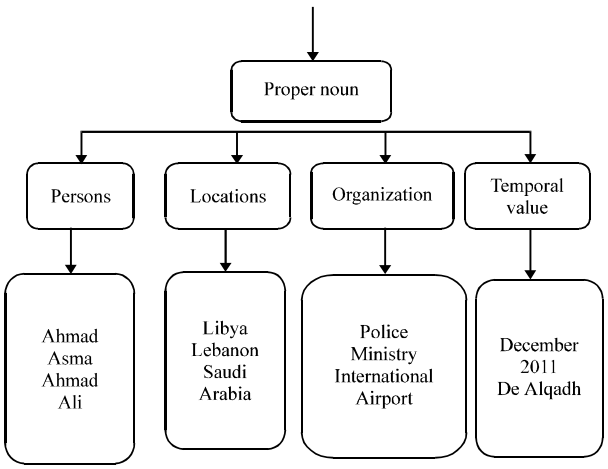

Fig. 1: Summarizes the types of proper noun

Table 1: Proper names with example

اسم العلم (NE)	Example
الاسم (Person)	محمد (Mohammed)
المكان (Location)	دبي (Dubai)
الشركة (Company)	مايكروسوفت (Microsoft)
التاريخ (Date)	السبت (Saturday)
الوقت (Time)	3:45 pm
السعر (Price)	450$
القياس (Measurement)	كيلومتر 40 (km)
رقم الهاتف (Phone No.)	0123768765

Table 2: Types of the information

Type	Arabic sentence	English translation	Proper noun
Noun	[illegible]	The union president Ahmed Ali	Ahmed Ali
Adjective	[illegible]	And the Iranian president Ahmed declared	Ahmed

Table 3: Type of sentences in NEs

Arabic sentences	English translation	Type of Arabic names	Type of English names	Type of proper noun
[illegible]	Hosni Mubarak met with the Israel former prime minister in Israel	[illegible]	Hosni Mubarak Kofi Annan	Person name
[illegible]	Hosni Mubarak and Kofi Aanan adopt a resolution decision between Lebanon	[illegible]	Lebanon and Israel	Location name
[illegible]	The Security Council will arrive to reach a resolution to the crisis next week.	[illegible]	the Security Council	Organization name
[illegible]	In October/December 2004 about 168	[illegible] 2004, 168	October December 2004, 168	Temporal value

Table 4: Accuracy for each class by using the rule-based system

Class	Precision (%)	Recall (%)	F-measure (%)
Person	100.0	70.6	82.76
Location	100.0	96.6	98.30
Organization	100.0	100.0	100.00
Temporal value	96.8	100.0	98.37
Overall F-scores	99.2	91.8	94.86

DESIGN ARCHITECTURE AND ALGORITHM

Figure 2 represents the whole process. Various methods of analysis and classification will be used before obtaining the final result. The methods which are chosen and emphasis depends largely on the design of the system, however, the system includes the following stages as shown in Fig. 2.

After input , any type of ANER sentences, the system includes three pre-processing modules that need to be used before the NER task.

The utilization of these modules depends on the nature of the input. These modules are used in the case when the input is raw text. The modules are segmentation, tokenization and POS tagging as follows.

The segmentation module: In this step the input text is segmented into several sentences. Besides, the boundaries of the sentence can be classified by symbols such as, end of line, punctuation and full stop. As a result of this segmentation the output will be annotations for each boundary as well as annotations for each sentence.

The Tokenizer: The tokenization is the process that analyses and splits the input text into a number of tokens such as, word, number, symbol, space, etc. The system will take the raw text from the collection of documents (corpus) and tokenize each text into tokens, so that they can be fed into a POS tagger for additional processing. The tokenizer is responsible for defining word boundaries,

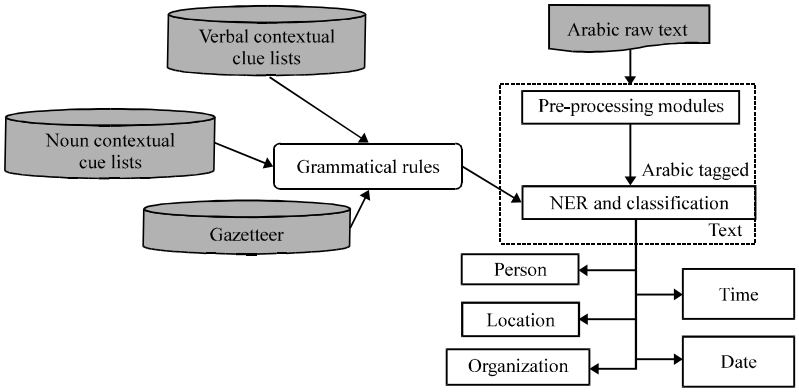

Fig. 2: ANER architecture

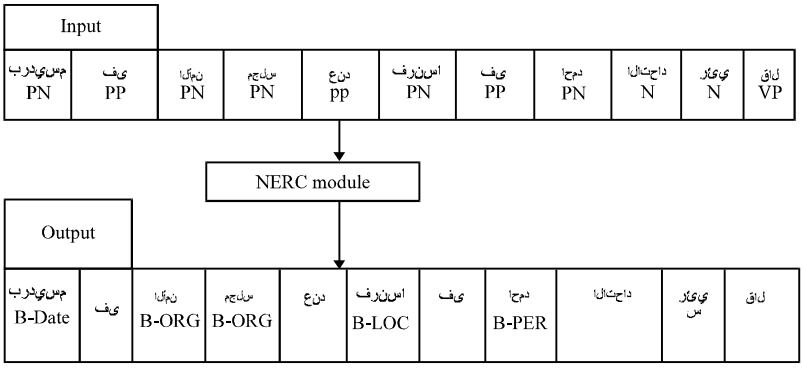

Fig. 3: The Input and the output of the NERC module

abbreviations and numbers. The Arabic token can be a number as a sequence of digits or a word as a sequence of connected letters or a conjunction or a symbol represented as,?, {, etc (Attia and Rashwan, 2004).

Arabic POS: POS tagging is the process of assigning a unique POS tag such as noun, pronoun, verb, adjective, or other tags to each word in a sentence. It reflects the word syntactic category based on its context for the purposes of resolving lexical ambiguity.

NE recognizer and classifier (NERC): The NERC module is the heart of this research (Fig. 3). As stated before, this module is designed using rule based approach. In fact, it consists of set of rules that uses sentence structure (i.e., here the POS of the word or words to be classified and the POS of its/their surrounding words).

Contextual clue lists (Verbs and Nouns): These lists contain special verbs or nouns that usually appear in the surrounding words (before or after) of NEs. Contextual Clue Lists are lists of keywords that can help identify some entities within documents. These lists are usually constructed by deep and careful study the NEs and their surrounding words in large documents.

Gazetteers: A set of lists that contains the names of well-known names such as people's names, organizations names, locations names, days of the week, etc. However, the NEs especially person names are countless.

Rules for person NE classification: First discussed how describe grammars for persons names work in general, then describe the grammars for organization. As we know, NEs may appear in any position within the sentence and can be next to a keyword of far away from it. Hence, we have identified a set of rules that recognize and classified ANEs. The rules here is depend on surrounding words, their POS tags, the word position and how far is it from a contextual clue words. In fact, the developments of contextual clue lists play a central role in the design of the rules. Words from the contextual clue lists are not candidate person names but are only used as keywords to find the position of person names in the text (Liu *et al*., 2010).

Rule 1: IF the current word and the next subsequent words (w_i.... .w_j) are proper nouns and they are followed by a Noun Contextual Clue Word or more for person names.

Then the current word and the next subsequent words (w_i... ... ... .w_j) are person names.

This rule means that if the current word and the next subsequent words from the input text matches any word of the Person Name list, or the POS of the current word and the next subsequent words are Proper Noun (PN) and the fourth word matches any word from the Noun Contextual Clue Word for person names and the fifth word matches any word from noun Contextual Clue Word for person names. Then the current word and subsequence words will be recognized as a person name. The output of this rule is the person name: (Omar Bin al-Khattab, عمر بن الخطاب).

Rules for organization NE classification:

Rule 2: IF (the combination of the POSs of (the current word and the immediately next words) is matched with the organization patterns) AND the current words are preceded immediately with a Verb Contextual Clue Word for an organization name.

Then the current word and its immediately next words is an organization name.

This rule means that if the first word from the input text matches any word from the Verb Contextual Clue Word for an organization name and the second and the third words match any word from the organization patterns or the POS of the second word is a Proper Noun (PN) and the POS of the third word is a Proper Noun (PN), then the second and third words will present the organization name.

The output of this rule is the organization: (Security Council, مجلس الامن).

RESULTS AND DISCUSSION

In general, the aim of this experiment was to investigate whether ANER system, sufficiently robust to be extracted from ANER. To evaluate the rule-based approach that has been used for system to recognize four types of NEs which are persons, locations, organizations and temporal values in the Arabic political domains. The experiment was applied to the set of test data include 35 sentences in Arabic NER in Political domain.

There are problems that appeared in information extraction, implementation of ANER system then classify the problems and assigns suitable scores for them. The score is given by human experts in extraction and it measures the differences between the human evaluation depending on the amounts of the magnitude of the error mismatch in structure.

The error mismatch: There are some mismatches of error test example which arise some problems in the output sentence. The following specifies the problems that appeared in the generation sentence:

- As have mentioned, The dada taken the political domain from some Arabic newswires. However, some of the persons' names are shown as short names (the first letter from the names) such as for The Allah or the name Faisal. This causes a problem because it cannot understand the meaning of names such as these
- Some of the person names do not appear because of the misprints such as due to forgetting to write a letter in the name or writing the name in the wrong way. Here, the rules cannot extract the person names in this case because they follow a particular specification depending on the rule structure
- We have used indicative noun words and verb words in some cases that are a help to recognize the names. However, sometimes the person names do not show because some of the indicative words come in the dual and plural
- Most of Arabic person names are NEs. However, sometimes it comes as nouns. In this case, the system may not show up these names

Evaluation settings: Used the standard evaluation measures in information extraction (Diab *et al*., 2004) i.e., (i.e., Precision, Recall and F-measures) have been used in the evaluation of the proposed method. Precision and recall are used as metrics for evaluating the correctness of the pattern recognition algorithm. Where precision is the percentage of NEs found by the system and which are correct. It can be expressed as:

$$\text{Precision} = \frac{\text{No.of correct NEs}}{\text{Entities found by algorithm}}$$

Also as: recall is the percentage of NEs existing in the corpus, it can be expressed as:

$$\text{Recall} = \frac{\text{No.of correct NEs}}{\text{Entities found by corpus}}$$

Recall and precision measures are combined in a single accuracy measure called the F-measure that is calculated as:

$$\text{F}-\text{measure} = \frac{2 \times \text{recall} \times \text{precision}}{\text{recall} + \text{precision}}$$

Experimental results: The goal of the experiments that are shown here is to evaluate the performance of the rule-based approach compared with another evaluation method. The results of our system were compared with the result of the human evaluation method for NER while using the same data set. Table 4 shows the accuracy in terms of the precision, recall and F-measure for each class of the NEs (person names, locations, organizations, temporal value) in the Arabic language by applying our system.

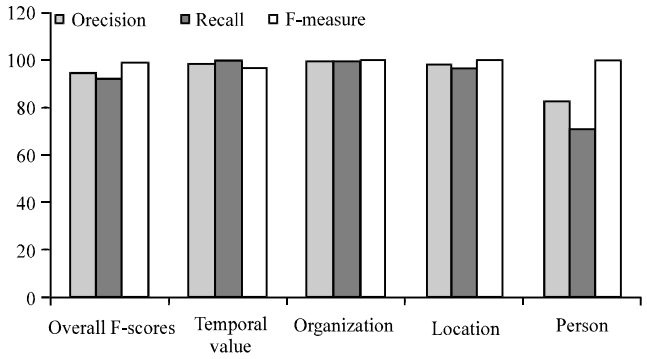

Fig. 4: The measurement by using rule based

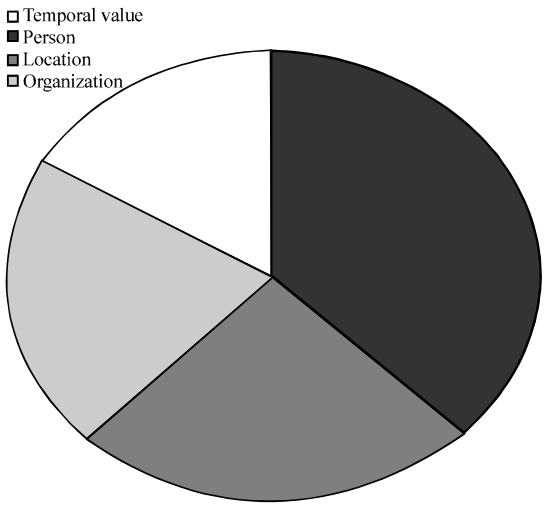

Fig. 5: The percentage of each named entities in the ANERcorp

Figure 4 shows the comparison between the three standard evaluation measures for classes of NE by applying the rule-based approach. The aim of this comparison was to show the differences in percentage of each measure in each class. It can be observed that the precision measure in persons, locations and temporal value and organizations classes gave a higher percentage than the other two measures.

Corpus: We have evaluated the system against humans, whereby we take the human annotated corpus and evaluate our system and then compare the results against this corpus (Diab *et al.*, 2004). The rule set was applied to all experiments to identify (person, location, organization and temporal value) and the output of the system was compared with a human standard set of the input text. In addition, the percentages of NE in the evaluation Corpus are: person 37.3%, location 25%, organization 21.2% and temporal value 16.5% as shown in Fig. 5, the percentages of NEs per class in the evaluation corpus (Benajiba *et al.*, 2008).

CONCLUSION

This study has been concentrated on issues in the implementation of a rule based ANER system which identifies each of the person names (PERS), location names (LOC), organization names

(ORG) and temporal value (TEMP) found for the Arabic language. A major goal of this system is to develop a model based on Rules to classify NE in Arabic text. We showed that the experiment is to investigate whether ANER system is sufficiently accurate for extraction Arabic NER. The output accuracy of our system is around 94.86% for extraction ANER in Political domain. We applied on 35 sentences in Arabic NER in Political domain and we faced some problems from mismatching between input and output. In future works solve the ambiguity and try different set of features and different rule types. These improvements will raise the correctness of the extraction from 94.86-100%.

REFERENCES

Abuleil, S., 2004. Extracting names from Arabic text for question-answering systems. Proceedings of the 7th International Conference of Computer-Assisted Information Retrieval Applications, April 26-28, 2004, Avignon, France, pp: 638-647.

Attia, M. and M.A.A. Rashwan, 2004. A large-scale Arabic POS tagger based on a compact Arabic POS tag set and application on the statistical inference of syntactic diacritics of Arabic text words. Proceedings of the Arabic Language Technologies and Resources International Conference, September, 2004, NEMLAR, Cairo, Egypt, pp: 1-6.

Benajiba, Y., M. Diab and P. Rosso, 2008. Arabic named entity recognition using optimized feature sets. Proceedings of the Conference on Empirical Methods in Natural Language Processing, October 25-27, 2008, Honolulu, HI., USA., pp: 284-293.

Diab, M., K. Hacioglu and D. Jurafsky, 2004. Automatic tagging of Arabic text: From raw text to base phrase chunks. Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics, May 2-7, 2004, Boston, MA., USA., pp: 149-152.

Kaiser, K. and S. Miksch, 2005. Information extraction: A survey. Vienna University of Technology, Institute of Software Technology and Interactive Systems, Vienna, Technical Report, Asgaard-TR-2005-6, May 2005.

Liu, B., L. Chiticariu, V. Chu, H.V. Jagadish and F. Reiss, 2010. Automatic rule refinement for information extraction. Proc. VLDB Endowment, 3: 588-597.

Moisiadis, F., R. Genrich, R. Stair and G. Reynolds, 2008. Principles of Information Systems. Thomson Course Technology, Boston, MA., USA., Pages: 481.

Nadeau, D. and S. Sekine, 2007. Survey of named entity recognition and classification. Ling. Invest., 30: 3-26.

Pouliquen, B., R. Steinberger, C. Ignat, I. Temnikova, A. Widiger, W. Zaghouani and J. Zizka, 2005. Multilingual person name recognition and transliteration. J. CORELA-Cognition Representation Lang., 2: 1638-5748.

Sekine, S., 2004. Named entity: History and future. Department of Computer Science, New York University, USA. http://www.cs.nyu.edu/~sekine/papers/NEsurvey200402.pdf.

Shaalan, K. and H. Raza, 2009. NERA: Named entity recognition for Arabic. J. Am. Soc. Inform. Sci. Technol., 60: 1652-1663.

Vergyri, D., K. Kirchhoff, K. Duh and A. Stolcke, 2004. Morphology-based language modeling for Arabic speech recognition. Proceedings of the International Conference on Spoken Language Processing, October 4-8, 2004, Jeju Island, Korea, pp: 2245-2248.

Zaghouani, W., B. Pouliquen, M. Ebrahim and R. Steinberger, 2010. Adapting a resource-light highly multilingual named entity recognition system to Arabic. Proceedings of the International Conference on Language Resources and Evaluation, May 17-23, 2010, Valletta, Malta, pp: 563-567.