International Journal of Engineering & Technology, 8 (1.11) (2019) 161-164

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

RANER: RDI Framework for Arabic Named Entity Recognition

Amr M. Sayed[1],Sherif Abdou[1], Mohsen Rashwan[2] , Hassanin Al-Barhamtoshy[3]

[1]Faculty of Computers & Information, [2]Faculty of Engineering
Cairo University, Egypt
[3] Faculty of Computing and Information Technology, King Abdulaziz University
Jeddah, Saudi Arabia
**Corresponding author E-mail: hassanin@kau.edu.sa*

Abstract

Named Entity Recognition (NER) task allows NLP applications to extract proper names from text. In addition, a significant component affects the performance of other NLP tasks such as text summarization, topic detection and key phrase extraction. Although many researches are conducted to enhance the NER task, only limited researches in Arabic named Entity Recognition have been performed. Researches in this field use either rule-based approach or machine learning approach. This paper introduces a solution for the NER problem using a machine learning approach, which combines Conditional Random Fields (CRF) classifier and predefined gazetteers. Our system uses syntactic features and morphological lookup-table features to train the classifier. This features extraction approach saves the time of morphological features that depend on analyzers without affecting the precision of the system. We evaluate our system by way of experiments using different sets of features to extract three of main named-entity types; Person, Location, and Organization. Experimental results showed that our approach has achieved better performance that rule-based approaches. Our system achieved a performance of 70.7%, 90.0%, and 79.2% in term of recall, precision, and F-measure, respectively..

Keywords: *named entitiy recognition; condetional random fields; machine learning-based approach; natural language processing.*

1. Introduction

A Named Entity (NE) is a textual phrase that clearly identifies one of pre-defined categories such as persons, organizations, and locations names. A phrase that belongs to a specific named entity class may belong to another class in different context. In addition, it may be a non-named entity in another context. NER seeks to identify these named entities within text through detecting the sequences of n-grams that belong to one of the pre-defined classes set. NER is a powerful task that improves the performance of NLP applications. Many researches stated that it’s better to collect NEs first, as the valuable information in text located at higher rates around NEs [1, 2].

In this paper, we focus on Arabic Named Entity Recognition (ANER). Arabic language is not limited to a minority. All the Arab peoples as well as the Islamic communities speak it. Arabic is a highly complex language on the levels of morphology and syntax [3]. A single word may have different meanings when it appears in different contexts. In addition, many words with different syntaxes may have the same meaning. Arabic language embeds five challenging issues for NLP tasks which make it more complex than English [4]. First, English NEs usually starts with capital letters, which make it easy to detect positions of NEs in text, unlike Arabic, which is not a case-sensitive language. Second, an Arabic word may have multiple prefixes and suffixes. This makes chunking significant to split an Arabic word in to stem and affixes. Third, Arabs may write the same word in different spelling deviations and typographic variants. Fourth, as mentioned above, single Arabic word may have different meanings within different contexts. Finally, Arabic resources are rare compared to English and collecting them is a very time consuming task.

We have adopted the machine-learning approach, in particular CRF, using syntactic and morphological features as well as NEs gazetteers to develop our system. We obtain the advantage of rule-based approaches by introducing these rules in the form of features during the training process to improve the performance of the classifier as well as decreasing their problems when they are used as a stand-alone rule-based component. Our approach has proven to be an adaptable model that can be easily modified to deal with new NEs. In addition, it can be simply modified to update the used feature set in the training process of the classifier. To evaluate our system we used the standard measures; precision, recall, and f-measure which show that the achieved performance is satisfactory.

The rest of this paper is organized as follows: Section 2 introduces a general overview of NER developed approaches especially in Arabic. Section 3 describes our NER problem statement. Section 4 describes the data preparation effort done. In addition, it shows the distributions of used data. Section 5 illustrates the architecture of out proposed system and its main components in details. Section 6 discusses the experiments that were processed to evaluate our approach as well as the effect of using different combinations of features. Finally, we draw some conclusions and comments regarding our future work in Section 7.

2. Related Work

Many researches focused on the ANER task in the past few years. Unlike English researches, which walk steadily through the NER task and reported the same machine learning approach (Maximum Entropy) to be the most successful language independent approach [5], ANER systems is still experiencing some confusion.

Many Arabic researches use the rule-based approach to develop NER systems. TAGARAB - an Arabic named entity recognizer that was developed by Maloney and Niv - uses morphological analysis to enhance the performance of a pattern-recognition engine [6]. The performance achieved by this engine in term of precision was 86.2% and 94.5% for Person and Location classes respectively. A rule-based technique was presented by Abuleil to extract three main types of named entities from text [7]. This technique use a set of rules generated through the detection of relationships between named entity phrases and their non-named entity counterparts in the text. This rule-based approach achieved precision of 90%, 93%, and 92% for Person, Location, and Organization classes respectively. A parallel corpora in Spanish and Arabic was used by Samy to tag Arabic names in text with the help of a Spanish NE tagger [8]. Arabic sentences should be translated into Spanish first. Then the Spanish NE tagger can detect the names that located in the Spanish translation. This model achieved precision of 84%.

Recently, most of researches focus on machine-learning approaches due to their superiority over rule-based approaches in terms of time consuming, flexibility, and domain independence. Benajiba et al., use n-grams and maximum Entropy to detect NEs in Arabic text [9]. This model achieved a precision of 82.17%, 61.54%, 45.16%, and 54.21% for Location, Miscellaneous, Organization, and Location classes respectively. Benajiba et al., use morphological features to train a Conditional Random Fields (CRF) classifier and got precision of 93%, 71%, 84.23%, and 80.41% for the same four classes respectively [10]. Authors in [11] use Support Vector Machines (SVM) classifier trained with lexical, contextual and morphological features. The SVM classifier got f-measure of 82.71% at testing on ACE 2003, Broadcast News genre.

Authors in [12, 13] use a combination of bidirectional Long Short-Term Memory (LSTM) and Conditional Random Field (CRF) to detect NEs in Arabic social media contents. They got f-measure of 71.52% and 85.71% respectively. Limsopatham and Collier [14] use bidirectional LSTM to automatically induce and leverage orthographic features for performing Named Entity Recognition in Twitter messages and got f-measure of 65.89%.

To sum up, according to the previous literature overview, Arabic Named Entity Recognition still needs more research and effort to build a mature enough system capable of dealing with Arabic texts in an accurate way. Most of NER techniques are not stable when they are applied on different domains. The domain of collected texts and their preparation is an effective phase during the preprocessing steps.
PROBLEM STATEMENT

We cast the Arabic Named Entity Recognition task as a sequence labelling problem. Each sentence is a sequence of words (tokens) which should automatically be assigned to different tags. A tag is an indicator of token's membership to a named entity category or class. Conditional Random Fields (CRF) is still one of the most powerful and effective techniques for the NER task. It achieved high performances in traditional and recent researches [4, 10, 12, 13]. In addition, CRF still able to withstand modern techniques such as deep learning and it can achieve results that compete with results of such modern techniques [15]. As previously stated, ANER is considered an important preprocessing tool to other NLP applications. Performance of NER task are clearly reflected on the performances of such applications. All of these points motivated us to enhance the performance of Named Entity Recognition task for Arabic Language by applying precise and powerful preprocessing techniques, which are the data preparation and features extraction and filtration, and then apply one of the best machine learning techniques (CRF) for the training and the preparation of the ANER model.

3. Data Preparation

To build a mature enough recognition model, an accurate tagged corpus is needed at the stages of training as well as testing. Table 1 shows the distributions of named entities on our corpora. Three datasets, tagged according to the IBO2 annotation (table 2), are used to build the training and testing named entities corpora:

ANERcorp: an Arabic corpus consists of more than 150,000 tokens [9].

AQMAR: an Arabic corpus consists of 28 articles with more than 74,000 tokens [15].

News Dataset: a manually collected 100 Modern Standard Arabic Documents from news (tagged manually).

Table I. Named Entities Distributions On Training And Testing Corpora

NE Class		Training Data		Testing Data	
Persons(B/I)		3,551	(3,551/2,669)	1,693 (1,693/1,266)	
Organizations(B/I)		1,823	(1,823/1,429)	854 (854/697)	
Locations(B/I)		4,296	(4,296/878)	1,795 (1,795/371)	
Others (not NE)		144,730		65,893	

Table II. Ibo2 Annotation For Named Entity Classes

Tag	Description
B-Pers	Beginning of person entity
I-Pers	Words inside person entity
B-Org	Beginning of organization entity
I-Org	Words inside organization entity
B-Loc	Beginning of location entity
I-Loc	Words inside location entity
O	Words that are not named entities

We use the ANERGazet gazetteers for person, location, and organization classes. We manually removed some entities that may cause ambiguity. For example, "نيمأ" may be used as a person name "Ameen" or a moral character "Honest". The three gazetteers, after adjustment, consist of 1488, 1517, and 346 entities for person, location, and organization classes respectively.

4. Proposed System Architecture

Machine learning-based approaches have their points of strength as well as points of weakness. To overcome the weaknesses arising from the use of machine learning-based systems, we apply some rule-based techniques in the features extraction phase. We used two types of features; word features and indicators features. Word features consist of seven different features (WF1:WF7); 1) word itself, 2) first two letters, 3) first three letters, 4) last two letters, 5) last three letters, 6) light stem of the word which is generated by a rule-based technique that removes the common affixes from the word, and 7) word form which is generating from a lookup table for one million frequent Arabic words, this lookup table is generated using RDI morphological analyzer [16]. Indicators features consist of five different features (IF1:IF5); 1) job name indicator that checks if the current word is a job name (e.g., “سيئرلا” which means president) using a manually collected set of 242 jobs names, 2) organization indicator that checks if the current word is a common prefix word for an organization (e.g., “ةكرش” which means company) using a manually collected set of 38 organizations prefixes, 3) in persons gazetteer indicator, 4) in location gazetteer indicator, and 5) in organization gazetteer indicator. Fig. 1 and fig. 2 show the architecture of the proposed system and its main components. They show training and testing phases components respectively.

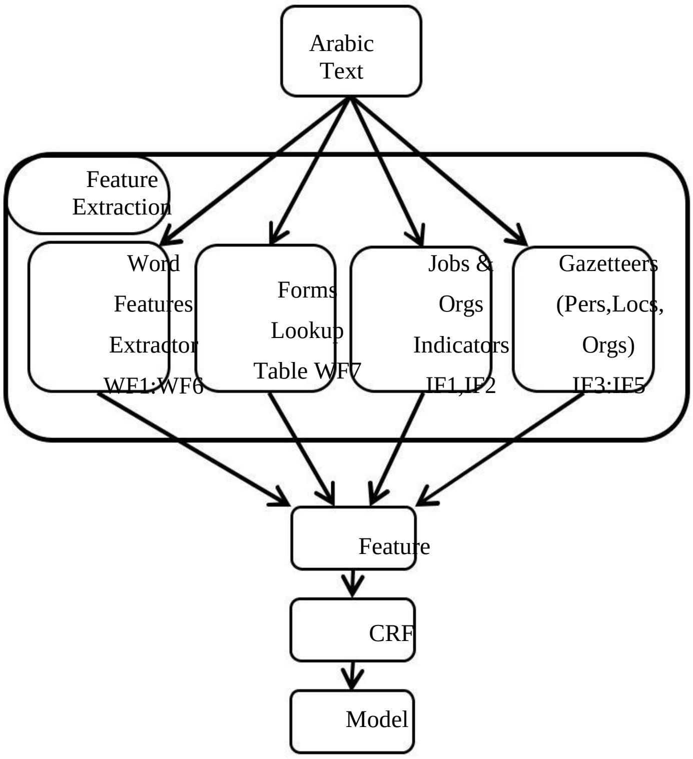
Arabic Text
Feature Extraction
Word Features Extractor WF1:WF6
Forms Lookup Table WF7
Jobs & Orgs Indicators IF1,IF2
Gazetteers (Pers,Locs, Orgs) IF3:IF5
Feature
CRF
Model

Fig. 1. Training Phase Components

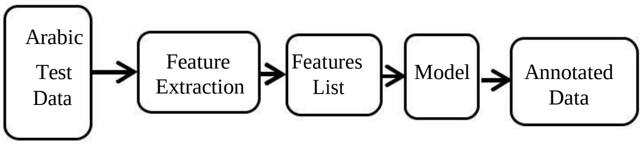

Fig. 2. Testing Phase Components

We used the CRF++ tool during the training and testing phases. In addition, we enable the classifier to take features of previous and next words in consideration as well as the current word features. Features combinations and their different effects on the system performance are shown in the next section.

5. Experiments and Discussion

In order to carry out some experiments we have trained and tested our system using different combinations of features to study their different effects on results in terms of recall(R), precision(P), and F-measure(F). Word features (WF1:WF7) affects the performance of the whole system in the named entity recognition task regardless of the type of this named entity. Indicators features (IF1:IF5) affects the performance of the system by affecting the performance of the classifier towards certain types of named entities. For example, IF1, which concerns about jobs names, affects the performance of the recognition of persons’ names directly.

Table III. Crf Results Using Different Combinations Of Features

Features	R	P	F
WF1	25.7%	87.8%	39.8%
WF1:WF5	46.3%	78.3%	58.2%
WF1:WF6	68.5%	**90.0%**	77.8%
WF1:WF7	69.7%	89.5%	78.4%
WF1:WF6, IF1	67.3%	89.6%	76.9%
WF1:WF6, IF1, IF2	69.1%	89.5%	78.0%
WF1:WF7, IF1, IF2	70.6%	89.6%	79.0%
WF1:WF7, IF1:IF3	70.4%	89.3%	78.7%
WF1:WF7, IF1, IF2, IF4	**70.7%**	**90.0%**	**79.2%**
WF1:WF7, IF1, IF2, IF5	70.0%	89.1%	78.4%
WF1:WF7, IF1:IF5	**70.7%**	89.9%	79.1%

Word features dramatically affect the results of the classifier. Using features WF2, WF3, WF4, and WF5 adds 19.4% to the F-measure that is generated by using WF1 only. Features WF2:WF7 have improved results as they reduce the “Out of Vocabulary” cases that may result from the use of the first word feature alone. Adding the feature WF6 to the previous combination of features adds another 19.6% to the F-measure. The use of persons and organizations gazetteers affects the results negatively as they add some ambiguities to the classifier.

The best results are achieved when all features are used except persons and organizations Indicators.

Table 4 shows the detailed results of our system using the best combination of features. Location class achieved the best results overall the remaining classes as locations names in text are straight forward and appear in common contextual patterns. Due to the lack of tagged organizations in the training corpus, this class achieved the worst results. In our training corpus, the number of organizations is not great enough if compared with persons and locations. We carried out a 10-fold cross validation to evaluate our system’s performance. Table 5 shows our system’s highest performance, in term of F-measure, compared with results of other systems when applied to ANERcorp dataset.

Table IV. Detailed Performance Of Our System Using Best Features Set

Class	R	P	F
LOC	**75.0%**	90.0%	81.8%
PERS	74.3%	**92.2%**	**82.3%**
ORG	57.8%	84.8%	68.8%
Total	**70.7%**	**90.0%**	**79.2%**

Table V. Results Of Other Systems Compared With Our System's Highest Performance

System	LOC	PERS	ORG
ANERsys 1.0	80.0%	46.0%	36.0%
ANERsys 2.0	86.0%	52.0%	46.0%
CRF-based [10]	**89.0%**	73.0%	65.0%
Hybrid CRF-based [4]	88.0%	68.0%	70.0%
Our System	**89.0%**	**86.0%**	**77.0%**

6. Conclusions and Future Work

In this paper, we propose an NER system for Arabic text that follows a machine learning-based approach. This system uses a CRF model with syntactic and lexical features. We present our experiments, which aim at selecting the best combination of features to improve the performance of ANER.

The results showed that using word features could obtain higher performance with respect to ANERsys. We used a lookup table of forms instead of morphological analyzers due to the high complexity of such systems. Experimental results on our mixed dataset have shown F-measure of 81.8%, 82.3%, and 68.8% for Location, Person, and Organization respectively. The results showed that we have obtained more improvement in precision than in recall.

In the next future, we plan to increase the size of our mixed corpus and to gather data from different domains in order to obtain better results. We plan to use a deep stemmer that follows a morphological analyzer instead of the light stemmer. We also

plan to carry out experiments using different feature-sets, and explore the possibility of using POS-tag. Furthermore, we plan to conduct a comparative study between SVM and CRF and to develop a hybrid system that integrates the many probabilistic models.

Acknowledgment

The National Plan funded this project for Science, Technology and Innovation (MAARIFAH) – King Abdulaziz City for Science and Technology -the Kingdom of Saudi Arabia– award number (11-INF-1997-03). The author thanks Science and Technology Unit, King Abdulaziz University for technical support. In addition, the author would like to thank the Research Lab of the engineering company for digital systems development (RDI). All authors thank profoundly RDI endless financial support through the research experimental phase.

References

[1] N. Chinchor, "Overview of MUC-7," in Proceedings of the Seventh Message Understanding Conference (MUC-7), 1998.

[2] S. Abuleil, "Extracting Names from Arabic Text for Question-Answering Systems," in Proceedings of Coupling approaches, coupling media and coupling languages for information retrieval (RIAO 2004), Avignon, France, 2004.

[3] A. Al-Sughaiyer and I. A. Al-Kharashi, "Arabic morphological analysis techniques: A comprehensive survey," Journal of the American Society for Information Science and Technology, pp. 189-213, 2004.

[4] S. AbdelRahman, M. Elarnaoty, M. Magdy and A. Fahmy, "Integrated machine learning techniques for Arabic named entity recognition," IJCSI, vol. 7, pp. 27-36, 2010.

[5] O. Bender, F. J. Och and H. Ney, "Maximum Entropy Models For Named Entity Recognition," in Proc. of CoNLL-2003, Edmonton, Canada, 2003.

[6] J. Maloney and M. Niv, "TAGARAB: A Fast, Accurate Arabic Name Recogniser Using High Precision Morphological Analysis," in Proceedings of the Workshop on Computational Approaches to Semitic Languages, Montreal, Canada, 1998.

[7] S. Abuleil, "Extracting Names from Arabic Text for Question-Answering Systems," in Proceedings of Coupling approaches, coupling media and coupling languages for information retrieval (RIAO 2004), Avignon, France, 2004.

[8] D. Samy, A. Moreno and J. Guirao, "A Proposal for an Arabic Named Entity Tagger Leveraging a Parallel Corpus," in International Conference RANLP, Borovets, Bulgaria.

[9] Y. Benajiba, P. Rosso and J. M. Bened´ıruiz, "Anersys:

[10] Anarabicnamedentity recognition system based on maximum entropy," in Computational Linguistics and Intelligent Text Processing, 2007.

[11] Y. Benajiba and P. Rosso, "Arabic named entity recognition using conditional random fields," in Proc. of Workshop on HLT & NLP within the Arabic World, LREC, 2008.

[12] Y. Benajiba, M. Diab and P. Rosso, "Arabic named entity recognition: An svmbased approach," in Proceedings of 2008 Arab International Conference on Information Technology (ACIT), 2008.

[13] Y. Shao, C. Hardmeier and J. Nivre, "Multilingual Named Entity Recognition using Hybrid Neural Networks," in The Sixth Swedish Language Technology Conference (SLTC), 2016.

[14] M. Gridach, "Character-Aware Neural Networks for Arabic Named Entity Recognition for Social Media," in WSSANLP 2016 , 2016.

[15] N. Limsopatham and N. Collier, "Bidirectional lstm for named entity recognition in twitter messages," in WNUT 20