A Novel Hybrid Approach to Arabic Named Entity Recognition

Mohamed A. Meselhi[1], Hitham M. Abo Bakr[1], Ibrahim Ziedan[1], and Khaled Shaalan[2]

[1]Derpartment of Computer and System Engineering; Faculty of Engineering,
Zagazig University, Egypt
mohatef@zu.edu.eg, {hithamab,iziedan}@yahoo.com
[2]The British University; Dubai, UAE
Khaled.Shaalan@buid.ac.ae

Abstract. Named Entity Recognition (NER) task is an essential preprocessing task for many Natural Language Processing (NLP) applications such as text summarization, document categorization, Information Retrieval, among others. NER systems follow either rule-based approach or machine learning approach. In this paper, we introduce a novel NER system for Arabic using a hybrid approach, which combines a rule-based approach and a machine learning approach in order to improve the performance of Arabic NER. The system is able to recognize three types of named entities, including Person, Location and Organization. Experimental results on ANERcorp dataset showed that our hybrid approach has achieved better performance than using the rule-based approach and the machine learning approach when they are processed separately. It also outperforms the state-of-the-art hybrid Arabic NER systems.

1 Introduction

Named entity recognition (NER) is still an important task for improving the quality of many NLP applications such as Information Retrieval, Machine Translation, and Question Answering [1]. NER seeks to identify the sequence of words in a document that can be classified under a predefined category of named entity such as Person, Organization, and Location names. Arabic is a highly inflected language, with a rich morphology and complex syntax [2]. Generally, the significance of Arabic worldwide is too obvious to enumerate. The language is spoken by Arab world, and Islamic countries and communities. In this paper we concentrate on NER for Arabic. We integrate a rule-based NER component, a reproduction of NERA [3], with a machine learning NER component, in particular SVM, in order to obtain the advantages of both approaches and decrease their problems. The rule-based component depends on a set of grammar rules. Whereas, the machine learning component depends on a set of features extracted from the annotated text. The extracted features include morphological features that have been determined by the Morphological Analysis and Disambiguation for Arabic (MADA) tool[1][4] and gazetteer features (list of predefined NEs).

[1] http://www1.ccls.columbia.edu/MADA/

X. Shi and Y. Chen (Eds.): CWMT 2014, CCIS 493, pp. 93–103, 2014.

We successfully could identify some recognition errors by simple grammar rules from the comparison between the results of the rule-based component and the results of the machine learning component.

The remainder of this paper is organized as follows. Section 2 introduces a background on NER. Section 3 describes the structure of our proposed hybrid system and its components. Experimental results are discussed in Section 4. In Section 5 we give some concluding remarks.

2 Background

Named Entity Recognition (NER) was first introduced in 1995 by the Message Understanding Conference (MUC-6)[2]. The named entity task is mainly defined as three subtasks: ENAMEX (for the Person, Location, and Organization), TIMEX (for Date and Time expressions), and NUMEX (for monetary amounts and percentages).

We focus on Arabic NER that has several challenges and characteristics:

- Lack of capital letters in the Arabic orthography: A named entity in Latin languages is usually distinguished by a capital letter at the word beginning. However, this is not the case in Arabic which makes the detection of NE in text based on the case of letters more difficult. Some efforts to overcome this problem have used lexical triggers that are derived from analyzing the NE's surrounding context [5] while some others have used the gloss feature of the English translation of the NE produced by the MADA tool [6].
- Complex Morphology: Arabic morphology is very complex because of the language agglutinative nature [7]. There are three types of agglutinative morphemes: stems, affixes and clitics. The primitive form of the word is the stem. Affix letters are attached to the stem which has three types: prefixes attached before the stem, suffixes attached after the stem, and circumfixes that surround the stem. Clitics are also attached to the stem after affixes. They play a syntactic role at the phase level. Clitics are either proclitics that precede the word or enclitics that follow the word. The conjunction "و" (wa[3], and) [8] and object pronoun "هن" (hn, they-3rdP-fem) are examples of proclitics and enclitics, respectively. A more general example is the word "وسيكتبونها" (wasayaktubuwnahA, and-they-will-write-it).
- Ambiguity: It is optional in modern Arabic texts to include diacritics which most often lead to ambiguous situations i.e. different meaning [9]. For example, consider the word علم Elm which if diacritized as عَلَمEalam it means the noun "flag" or if diacritized as عِلمEilm it means the noun "science". In addition to the optional diacritization problem, Arabic words can differ in meanings depending on the context in which it appears. Consider the following two sentences: قالت جريدة الشرق الأوسط إن الأسد باق في منصبه (qAlt jrydp Al$rq Al>wsT <n Al>sd bAq fy mnSbh, 'Asharq Al Awsat said that al-Assad will remain in his position') and الإرهاب قضية مهمة في منطقة

[2] http://cs.nyu.edu/cs/faculty/grishman/muc6.html
[3] Habash-Soudi-Buckwalter transliteration scheme

الشرق الأوسط (Al<rhAb qDyp mhmp fy mnTqp Al$rq Al>wsT, 'Terrorism is an important issue in the Middle East'). The named entity, الشرق الأوسط (Al$rq Al>wsT, Middle East) might represents either an Organization or a Location name which can be resolved from the context.

- Lack of Resources: The lack of Arabic NER freely available resources or the expense of creating or licensing these important Arabic NER resources make NER task far more challenging. So, researchers had to build their own resources. We used ANERcorp[4] Corpus that developed by [10] for both training and testing. The ANERcorp includes 4901 sentences with 150286 annotated words for the NER task. The total number of named entities is 12989 tokens.

The survey by Shaalan (2014) presents background and the progress made in Arabic NER research.

3 The Approach

Both machine learning-based and rule-based approaches have their own strengths and weaknesses and by combining them in one system they could achieve a better performance than applying each of them separately. To the best of our knowledge, the latest Arabic NER system has adopted the hybrid approach was presented by Oudah and Shaalan [11]. The system consists of two dependent components including rule-based component and machine learning component where the output of rule-based component that represented as features file is passed as an input to machine learning component . Thus, any classification error in rule-based component will be trained by the classifier which leads to an error in the whole system. In addition, any modification in the grammar rules or the gazetteers requires retraining the classification model again. Our novel hybrid approach aims to fix the previous problems with fully independent components. It consists of three main components: machine learning component, rule-based component, and tag selection and correction component.

3.1 The Machine Learning Component

Arabic NER is challenging due to the highly ambiguous nature of Arabic named entities (NEs). A NE can very well appear as a non-NE in Arabic text. Moreover, the modern style of Arabic writings allows for optional diacritization and different methods for performing transliterations. So, these peculiarities of the language raise the need for Machine Learning (ML) algorithms which lend itself to the Arabic NER task. These algorithms usually involve a selected set of features, extracted from datasets annotated with NEs, which is used to generate a statistical model for NE prediction. In the literature, the ML approaches used in NER systems are Maximum Entropy (ME) [12], Hidden Markov Model (HMM) [13], Conditional Random Fields (CRFs) [14], [15] and Support Vector Machines (SVM) [16].

[4] http://www1.ccls.columbia.edu/~ybenajiba/

The machine learning component uses SVM because of its robustness to noise and ability to deal with a large number of features effectively [17]. SVM is a supervised machine learning algorithm which is based on Neural Networks [16]. SVM learns to find a linear hyperplane which divides the elements (features) in space into positive and negative classes with maximal margin. We used YamCha toolkit[5] that converts the NER task to a text chunking task.

In Machine Learning-based NER approaches, there are two phases: training phase and test phase, as illustrated in Fig. 1. The first phase generates the classifier (model) by using a set of classification features. In the second phase, the classifier generated by the training phase is utilized to predict a class for each token (word).

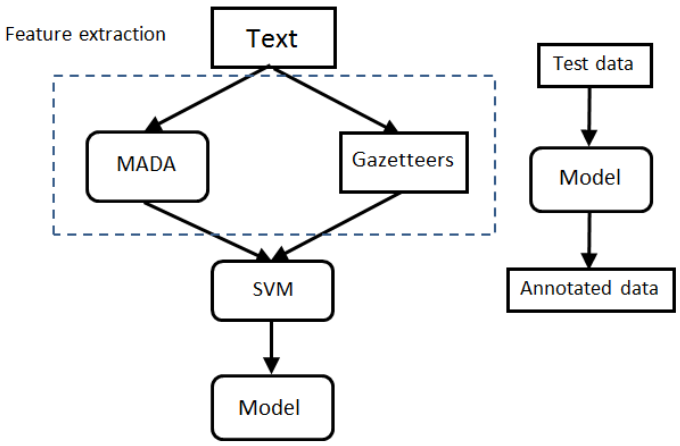

Fig. 1. Training and test phases

In the training phase each word is represented by a set of features and its actual NE's type in order to produce an SVM model that predicts the NE type. The selection of the subset to be utilized by a classifier is a very critical to the NER system's performance such that when optimized it can enhance the quality of the system dramatically. So, the first step in our hybrid approach is to extract the significant features from the training dataset. Then, we study the impact of each feature individually by adopting only one feature at a time and measure the system's performance in terms of F-measure metric. Finally, according to the performance achieved, we determine the optimized feature set for the proposed hybrid Arabic NER system.

3.2 The Feature Space

Feature selection refers to the task of identifying a useful subset of features chosen to represent elements of a larger set (i.e., the feature space). In the rest of this section, we discuss the feature space.

[5] http://chasen.org/•taku/software/yamcha/

- The word: It refers to the distribution of each NE type in the ANERcorp dataset. As shown in Table 1, the highest frequency is the Location NE. So, for example, the classifier would mostly recognize the sequence which consists of the word (الشرق, Al$rq) that is followed by the word الأوسط (Al>wsT, Middle) as a Location (East) name rather than an Organization (newspaper).

Table 1. Number of times each type was assigned to the word "الشرق" in ANERcorp

NE Type	Frequency
O	17
Person	0
Location	51
Organization	14
Total	82

- Contextual word feature (CXT): The features of a sliding window comprising a word n-gram that includes the candidate word, along with preceding and succeeding words For example, in the training corpus the verb "حضر" (HDr, attend) appears frequently before an NE of type Person. As such, the classifier will use this information to predict a Person NE after this verb.

- Gazetteer features (GAZ): A binary feature indicating the existence of the word in an individual gazetteer. In our hybrid system there are three gazetteers:
  - Location Gazetteer: countries, cities, rivers and mountains, etc.
  - Person Gazetteer: names of people.
  - Organizations Gazetteer: companies and other organizations.

- Morphological features (MORPH): A set of morphological information determined by MADA. We cover the following morphological features:
  - Aspect: One of the three aspects of the Arabic verb: perfective (ماضي, mADy), imperfective (مضارع, mDArE), or imperative (أمر, >mr).
  - Case: One of three cases of Arabic nominals: nominative (مرفوع, mrfwE), accusative (منصوب, mnSwb), or genitive (مجرور, mjrwr).
  - Gender: A binary value indicating the gender of the word, i.e. masculine or feminine.
  - Number: Any of the three values indicating the number of the word, i.e. singular, dual, or plural.
  - Mood: Any of the three Arabic moods that only vary for the imperfective verb: indicative (مرفوع, mrfwE), subjunctive (منصوب, mnSwb), or jussive (مجزوم, mjzwm).
  - State: Any of the three values: definite, indefinite, or construct.
  - Voice: A binary value indicating either passive or active voice.
  - Proclitics and Enclitics: exact clitics that are attached to the stem.

- Part-of-Speech (POS): A binary value indicating whether or not the POS tag (extracted by MADA) is a noun or proper noun.
- Gloss: A binary value indicating whether the English translation (gloss) provided by MADA starts with a capital letter.
- Lexical features (LEX): This feature considers the orthography of each token in the text. The usefulness of lexical features mostly appears when the same NE occurs in different places of the text but with some difference in the orthography. For example, in the sentence: أوباما يدافع عن برامج مراقبة الهواتف والإنترنت (>wbAmA ydAfE En brAmj mrAqbp AlhwAtf wAl<ntrnt, Obama defends program of surveillance phones and the Internet). Transliteration variant of foreign names is a common problem in Arabic writing. For example, Obama may be transliterated to Arabic as (أوباما, >wbAmA) or (أباما, >bAmA). So, in the training corpus, if Obama has only appeared with the first transliteration, the classifier cannot classify the second transliteration. However, when we used the last characters of the word as a feature, it would help the classifier to classify second transliteration also.

3.3 The Rule-Based Component

The initial version of the rule-based component is developed using the design documents of NERA, a NER system for Arabic [5] that is implemented using GATE[6]. Fig. 2 shows the construction of the Rule-based component. The system applies two procedures: firstly, recognizing and classifying NEs in text by the exact matching with gazetteers entries in the corresponding Person, Location and Organization Gazetteers. A sample of these gazetteers is shown in Table 2.

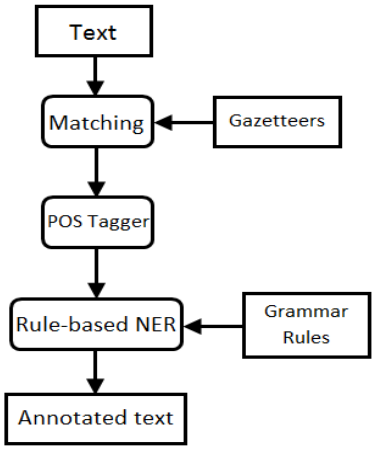

Fig. 2. Rule-based component

[6] http://gate.ac.uk/

Table 2. Sample entries in the three gazetteers

Complete Names	حسني مبارك
	Hsny mbArk
	Hosni Mubarak
City Names	الاسكندرية
	AlAskndryp
	Alexandria
Parties Names	الحزب الوطني الديمقراطي
	AlHzb AlwTny AldymqrATy
	National Democratic Party

Secondly, execute a finite-state transducer, based on a set of local grammar rules that are implemented using JAPE. The following example illustrates an implemented rule for recognizing an organization name. This rule identifies an organization name (token) that is preceded by a verb indicating an organization name (e.g."فاز", win) and followed by a nationality (e.g." الأردنية", Jordanian).

```
Rule: ORG1
Priority:30
(
{Lookup.majorType=="verb-org"}
({Token}):Org1
({DAL})
)
-->
:Org1.Organization = {rule="ORG1"}
```

Fig. 3. A rule implemented in JAPE for recognizing organization name using surrounding indicators

JAPE rules depend also on Stanford POS Tagger[7] that assigns part of speech category to each word in the text. Often proper noun tags mark the existence of NEs in the text. So, based on this available information, we modified the preceding rule to verify whether or not the targeted token is a proper noun.

[7] http://nlp.stanford.edu/software/tagger.shtml

```
Rule: ORG1_modified
Priority:30
(
{Lookup.majorType=="verb-org"}
({Token,Token.category ==
noun_prop}):Org1
({DAL})
)
-->
:Org1.Organization = {rule="ORG1"}
```

Fig. 4. A rule implemented in JAPE for recognizing organization name using its POS category

3.4 Tag Selection and Correction Component

Fig. 5 shows the architecture of our proposed Arabic hybrid NER system. The input data is applied in parallel to the rule-based and machine learning-based Arabic NER components. The tagging results for each component are compared with each other in order to agree on the final tags. The role of this component is to fine-tune the machine learning system's output by checking the most false negatives (i.e., missing annotations) and applying the correction using the tagging decisions determined by the rule-based component.

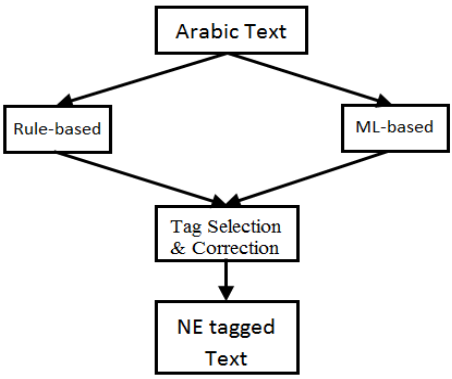

Fig. 5. Architecture of the hybrid System

4 Experiments and Results

In order to study the impact of the NER approaches we have examined each component separately before combining the results. In other words, we are dealing with

three annotated outputs from the machine learning component, rule-based component, and the overall hybrid output, respectively. The experimental setting for the machine learning component uses splits of the ANERcorp dataset into three datasets: 90% as a training dataset, 5% as a development dataset, and 5% as a test dataset.

As far as the machine learning component is concerned, the experiment proceeds in three stages. In the first stage, the training is applied on the training dataset using selected feature set and the results are analyzed to determine the best feature set. In the second stage, the training is applied on the combined training and development datasets using the best selected feature set. In the third stage, the classifier is applied on the test dataset and the results are reported and discussed.

The baseline feature set consists of tokens at window size that ranges from -1/+1 to -4/+4.We found that a context size of one previous token and one subsequent token (i.e. window size is 3) achieves the best performance in this task. The baseline model has achieved a precision of 91.86% and recall of 48.78%. This indicates that adding extra features would improve the performance by increasing its coverage.

As far as the rule-based component is concerned, the untagged version of the reference dataset (i.e. ANERcorp) is entered to GATE for processing, where the annotated result can be automatically evaluated using Annotation Diff tool of GATE which has an elegant GUI for presenting the results. Three results were obtained from machine learning, rule-based and hybrid approaches in terms of the standard evaluation metric, i.e. Precision, Recall and F-measures [18], for Person, Organization and Location, as presented in Table 3. This table shows that the results from machine learning approach (SVM) are better than that from rule-based approach whereas the hybrid system performs the best. Table 4 shows that our hybrid system outperforms the state-of-the-art Arabic hybrid NER system by [11].

Table 3. performance of each component in the system

Approach	Type	P	R	F
Machine Learning	**PER**	97.67	93.33	95.45
	ORG	89.19	88.00	88.59
	LOC	96.12	88.84	92.34
Rule-based	**PER**	97.98	90.00	93.82
	ORG	81.29	92.67	86.60
	LOC	95.26	88.05	91.51
Hybrid	**PER**	97.01	96.30	96.65
	ORG	90.00	96.00	92.90
	LOC	95.18	94.42	94.80

Table 4. Comparing best results of (Oudahand Shaalan, 2012) with our best results

	PER	ORG	LOC
Oudah and Shaalan, 2012	94.4	88. 2	90.1
Our approach	96.65	92.9	94.8

5 Conclusions

Our proposed hybrid NER approach integrates the rule-based approach with the ML-based approach in order to optimize overall performance. The two components responsible for the integrated approached are processed in parallel. A tag selection and correction component is used in order to fine-tune the machine learning system's output by checking the most false negatives (i.e., missing annotations) and applying the correction using the tagging decisions determined by the rule-based component. Experimental results on ANERcorp dataset, with F-measure have shown 96.65%, 92.9%, and 94.8% for Person, Organization, and Location, respectively. Therefore, our hybrid system outperforms the state-of-the-art of the Arabic hybrid NER system. Our study on the impact of the features indicates that when window size is 3 it achieves the best performance. For future work, the authors would like to increase the capability of the system in identifying other types of named entities. We are also considering the possibility of investigating different machine learning techniques other than SVM and study their impact on the overall performance of the hybrid NER system.

References

1. Benajiba, Y., Diab, M., Rosso, P.: Arabic Named Entity Recognition using Optimized Feature Sets. In: Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing, Honolulu, pp. 284–293 (2008)
2. Al-Sughaiyer, I.A., Al-Kharashi, I.A.: Arabic morphological analysis techniques: a comprehensive survey. Journal of the American Society for Information Science and Technology 55(2004), 189–213 (2004)
3. Shaalan, K., Raza, H.: NERA: Named entity recognition for arabic. Journal of the American Society for Information Science and Technology, 1652–1663 (2009)
4. Habash, N., Rambow, O., Roth, R.: MADA+TOKAN: A Toolkit for Arabic Tokenization, Diacritization, Morphological Disambiguation, POS Tagging, Stemming and Lemmatization. In: Proceedings of MEDAR, Cairo, Egypt, pp. 102–109 (2009)
5. Shaalan, K., Raza, H.: Arabic Named Entity Recognition from Diverse Text Types. In: Nordström, B., Ranta, A. (eds.) GoTAL 2008. LNCS (LNAI), vol. 5221, pp. 440–451. Springer, Heidelberg (2008)
6. Farber, B., Freitag, D., Habash, N., Rambow, O.: Improving NER in Arabic Using a Morphological Tagger. In: Proceedings of LREC 2008 (2008)
7. Habash, N.Y.: Introduction to Arabic Natural Language Processing. Mogran & Claypool Publisher (2010)
8. Habash, N., Soudi, A., Buckwalter, T.: On Arabic transliteration. In: Arabic Computational Morphology: Knowledge-based and Empirical Methods. Kluwer/Springer (2007)

9. Shaalan, K.: A Survey of Arabic Named Entity Recognition and Classification. Computational Linguistics 40(2), 469–510 (2014)
10. Benajiba, Y., Rosso, P., BenedíRuiz, J.M.: ANERsys: An Arabic Named Entity Recognition System Based on Maximum Entropy. In: Gelbukh, A. (ed.) CICLing 2007. LNCS, vol. 4394, pp. 143–153. Springer, Heidelberg (2007)
11. Oudah, M., Shaalan, K.: A pipeline Arabic named entity recognition using a hybrid approach. In: Proceedings of the 24th International Conference on Computational Linguistics, COLING 2012, India, pp. 2159–2176 (2012)
12. Borthwick, A.: A Maximum Entropy Approach to Named Entity Recognition Ph.D. thesis, Computer Science Department, New York University (1999)
13. Bikel, D.M., Schwartz, R.L., Weischedel, R.M.: An Algorithm that Learns What's in a Name. Machine Learning 34(1-3), 211–231 (1999)
14. Lafferty, J.D., McCallum, A., Pereira, F.C.N.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: Proceedings of the Eighteenth International Conference on Machine Learning, pp. 282–289 (2001)
15. McCallum, A., Li, W.: Early Results for Named Entity Recognition with Conditional Random Fields, Feature Induction and Web-Enhanced Lexicons. In: Proceedings of Seventh Conference on Natural Language Learning, CoNLL 2003 (2003)
16. Vapnik, V.: The Nature of Statistical Learning Theory. Springer, Heidelberg (1995)
17. Benajiba, Y., Diab, M., Rosso, P.: Arabic named entity recognition: An svm-based approach. In: The International Arab Conference on Information Technology, ACIT 2008 (2008)
18. Sitter, A.D., Calders, T., Daelemans, W.: A Formal Framework for Evaluation of Information Extraction, University of Antwerp, Dept. of Mathematics and Computer Science, Technical Report, TR 2004-0 (2004)