

ANERsys 2.0: Conquering the NER Task for the Arabic Language by Combining the Maximum Entropy with POS-tag Information

Yassine Benajiba and Paolo Rosso

Dpto. Sistemas Informáticos y Computación (DSIC),
Universidad Politécnica de Valencia, Spain
{ybenajiba, proso}@dsic.upv.es

Abstract. In this paper we describe an improved version of ANERsys, an Arabic Named Entity Recognition system for open-domain texts. The first version of ANERsys was totally based on the Maximum Entropy approach and was trained and tested with corpora which we have built ourselves. The results showed that the Maximum Entropy is an appropriate method to identify Named Entities in Arabic texts. However, in order to reach higher performance a greater effort needed to be done to improve the recognition of long proper names. Therefore, in the second version of ANERsys, we use a Part Of Speech tagger and a two-steps approach to enhance the performance of the system. Furthermore, we have used our own (now freely available on our website) corpora (ANERcorp) and gazetteers (ANERgazet) to train and evaluate ANERsys 2.0. We carried out several experiments to evaluate the performance of the system and to compare it with the online freely available demo version of the commercial system Siraj (Sakhr). The results show that the accuracy of the new version is 10 points above the old one and more than 7 points above the Siraj (Sakhr) system accuracy.

1 Introduction

In the sixth Message Understanding Conference (MUC-6)¹ the Named Entity Recognition (NER) task was defined as three subtasks: ENAMEX (for the proper names), TIMEX (for temporal expressions) and NUMEX (for numeric expression). The first sub-task is the one we are concerned about. ENAMEX was defined as the extraction of proper names and the classification of each one of them as: (i) Organization (named corporate, governmental, or other organizational entity); (ii) Location (name of politically or geographically defined location) or (iii) Person (named person or family).

Nowadays, we can only find in the literature two NER systems for Arabic which make impossible to carry out a comparative study and determine the best approach for the Arabic NER task. The first of these works [1] is totally based on the use of a set of keywords as triggers and a set of rules to extract the

¹ <http://cs.nyu.edu/cs/faculty/grishman/muc6.html>

proper names. The second one [2] uses a high precision morphological analysis. However, the shared task of CONLL 2002² and CONLL 2003³ was completely devoted to the language-independent NER task. The test-set in CONLL 2002 consisted of Spanish and Dutch corpora. In CONLL 2003 they used English and German corpora. Most of the best participations were ME-based systems [3][4][6][7]. In [8] a comparison was made between the Hidden Markov Models (F-measure 31.87) and the Maximum Entropy (ME) (55.77) approaches (when a more sophisticated set of features was used the ME approach reached an F-measure of 85.61). It is important to point out that in CONLL 2002 the corpora contained two columns: a first column for the words and a second one for the named entity tag, whereas in CONLL 2003 the corpora contained four columns: the first column for the words, the second one for the Part Of Speech (POS) tag, the third one for the syntactic chunk tag and finally the fourth one for the named entity tag. The best accuracy in CONLL 2003 was 88,76 [9] which is 7 points above the best accuracy in CONLL 2002 [5]. Thus, from the above study of the different systems we found out that the technique adopted by mainly the best participations is the ME.

The rest of this paper is structured as follows. In the second section we will emphasize the Arabic language characteristics which are related to the topic of the paper. Section Three will explain with details our approach to enhance the last version of the system. Section Four is dedicated to describe the data set we used to train and evaluate our system. Finally, in the fifth section we present the different experiments we carried out with ANERsys, whereas in the sixth section we draw our conclusions and future works.

2 Peculiarities and Challenges of the Arabic Named Entity Recognition Task

Some of the characteristics of the Arabic language make the NER task even more challenging:

(i) The Arabic language has a very complex morphology. Similarly to all the other Semitic languages, words are formed by inserting affixes to the root. Figure 1 shows a simple example of the composition of an Arabic word. This concatenative strategy to form the words causes data sparseness. Therefore, a very large training corpus is necessary for a good training.

We can find generally in the literature two possible solutions to overcome this obstacle. The first solution is to perform a light-stemming and consists of omitting all the affixes and keeping only the root morpheme of a word. Unfortunately, this solution is inappropriate for the NER task because it omits the prepositions which appear as affixes. Consequently valuable contextual information is lost. The other solution is to perform a text-segmentation. It consists of

² <http://www.cnts.ua.ac.be/conll2002/ner/>

³ <http://www.cnts.ua.ac.be/conll2003/ner/>

detaching the constituent morphemes of each word and separate them by the space character. Thus, the text-segmentation technique is able to decrease the data sparseness without causing any contextual information loss and, therefore, it is more convenient for the NER task.

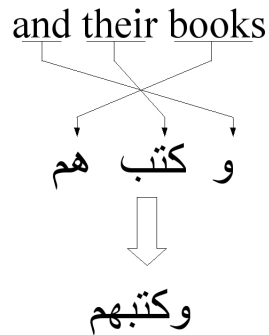


Fig. 1. A simple example of the composition of an Arabic word.

In the first version of ANERsys we used an heuristic which takes into consideration only prefixes. However, in the enhanced version of the system which we present in this paper we used a text-segmentation tool (see Section Four for more details);

(ii) The Arabic language does not support capital letters. This characteristic contributes significantly to harden the NER task for the Arabic language because it is an essential feature for the NER systems.

3 Lexical Resources for NER

As we have mentioned before, there is no freely available Arabic corpora for the NER task. For this reason we built our own corpora. We made the developed resources which we present in this section available on our web site to ease the further research activity of the Arabic NER task. Following we present some of the characteristics of our resources:

3.1 ANERcop: Corpora for Training and Test

ANERcorp⁴ was annotated in order to have exactly the same format of the CONLL 2002 corpora that we have mentioned above in the introduction. We have

⁴ <http://www.dsic.upv.es/~ybenajiba>

used the same classes defined in the MUC-6 (Person, Location and Organization) with an extra Named Entity (NE) class (Miscellaneous) for NEs which do not belong to any of the other classes. We have found the IOB2 scheme as the most appropriate one for our needs. This scheme considers that a word may either belong to one of the four classes mentioned above or belong to the class O. In case it is not a word belonging to the O class then it is either at the Beginning of a NE or Inside of a NE. Thus, any word of the text should be annotated as one of the following nine tags (Figure 2 shows an extract from ANERcorp) :

B-PERS, I-PERS, B-LOC, I-LOC, B-ORG, I-ORG, B-MISC, I-MISC, O.

ANERcorp consists of more than 300 articles (more than 150,000 tokens) manually collected from different sources and different kind types of articles. The annotation was done manually by one person to guarantee the coherence of the corpora. 11% of ANERcorp words are proper names. Their distribution along the different NE classes is shown in Table 1.

B-PERS	محمود
I-PERS	عباس
O	سيقرر
O	فور
O	عودته
O	من
B-LOC	الأردن
O	نوع
O	الخطوات
O	التي
O	سيأخذها
O	لأنهاء
O	الازمة
O	.

Fig. 2. An extract from ANERcorp

Table 1: Ratio of NE's by classes

Class	Ratio
PERSon	39%
LOCation	30.4%
ORGanization	20.6%
MISCellaneous class	10%

3.2 ANERgazet : Corpora for Training and Test

ANERgazet⁵ gathers three different manually built gazetteers:

(i) Location Gazetteer: 1,950 names of continents, countries, cities, mountains, etc. which were extracted from the Arabic Wikipedia⁶;

(ii) Person Gazetteer: 2,309 tokens of Arabic and non-Arabic names obtained from different web sources;

(iii) Organization Gazetteer: 262 names of international organization, soccer teams, etc. obtained from different web sources as well.

A more detailed description of the resources is available in the paper describing the first version of ANERsys [10].

4 Combining Maximum Entropy with POS Tag Information

The first version of our system [10] was completely ME-based where the contextual information (usually called features) of a word W_i was the only type of information the system relied on to determine its class C_i . Thus, the system's architecture was quite simple. In the training phase the system computes the weights of each feature with each of the classes, and in the second phase (i.e. the test phase), the system uses the weights that were computed earlier to classify each of the test corpus words (we draw the results of the first version of our system in Section Five). The results of our preliminary experiments showed that the ME is a convenient approach for the Arabic NER task. However, the system needed an improvement for a better recognition of the long NE's. Therefore, in the second version we have adopted a two-steps approach. The first step extracts the boundaries of the NE's. The second one classifies each of the NE's delimited in the previous phase (see Figure 3). Subsections 4.1 and 4.2 present further details on this enhanced approach.

4.1 Step 1 -Named Entities Boundaries Detection

As we mentioned above, the first step of our system concerns only the delimitation of the NE's. The input file to this first step should be an IOB2 annotated corpus. The delimitation of the boundaries is made initially by a ME-based and a POS-tag-based modules. Thereafter, the results are combined in a module which was placed at their outputs. Following we present a brief description of the mentioned modules: (i) The ME-based module uses an exponential model which can be illustrated by the following equation:

$$p(c|x) = \frac{1}{Z(x)} * \exp\left(\sum_i \lambda_i \cdot f_i(x, c)\right) \quad (1)$$

⁵ <http://www.dsic.upv.es/~ybenajiba>

⁶ <http://ar.wikipedia.org>

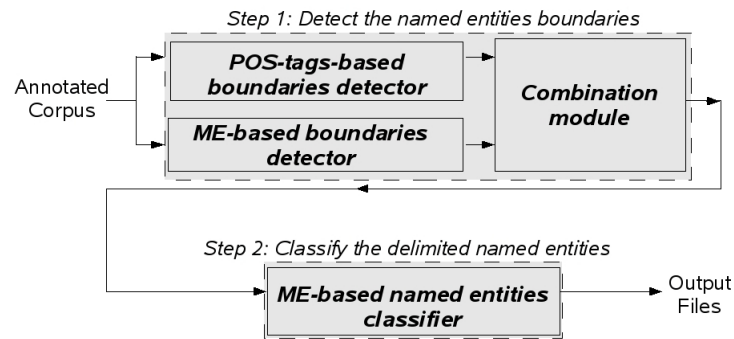


Fig. 3. Generic architecture of ANERsys 2.0

Where c is the class, x is a context information and $f_i(x, c)$ is the i -th feature. The features are binary functions indicating how the different classes are related to one or many classes. The λ_i weights are trained using only features related to the beginning and the inside of the NE's and $Z(x)$ is for normalization and may be expressed as:

$$Z(x) = \sum_{c'} \exp\left(\sum_i \lambda_i \cdot f_i(x, c')\right) \quad (2)$$

(ii) On the other hand, for POS-tag-based boundaries detection we have used Mona Diab's Arabic POS tagger. This POS tagger is freely available on her web site⁷ in a package together with a tokenizer and a Base Phrase (BP) chunker [11]. The author reports that all the tools of the package were trained on data derived from the Arabic Treebank. The model files are included in the package, hence the use of the mentioned tools does not require any type of annotated corpora. These tools were tested on a 400 Arabic sentences and the reported accuracies are very high. To delimit the boundaries we first select the phrases defined as Noun Phrases (tagged as NP) by the BP chunker. The following step is to keep only the NP's whose words were tagged as singular, dual or plural proper nouns (tagged as NNP or NNPS by the POS tagger). (iii) Finally, the combination module first conducts a union of the results of the previous two modules. Additionally, a second operation is performed to change the tags which were wrongly put as B-x instead of I-x or vice versa.

4.2 Step 2 Delimited NE's Classification

The second step of our approach is totally based on ME. We have similarly used the exponential model. The purpose in the second step is to classify each of the NE's delimited in the previous step as one of the four classes mentioned above in Section Three.

⁷ <http://www1.cs.columbia.edu/~mdiab>

4.3 Features Selection

The features used for both the ME-based modules of the system (the ME-based boundaries detector in the first step and the NEs classifier in the second step) are binary features. In order to determine an optimal features-set we have conducted several experiments with different sets. Moreover, our conclusion is that the following features-sets give the best results: -The ME-based boundaries detector module relies mostly on contextual information. Thus, the features-set is composed of the frequently preceding n-grams⁸, which is the compilation of a list of n-grams which were observed more than m times (The best results were achieved for m=25) preceding a NE in the training corpus. -Some of the ME-based NEs classifier features-set elements are: (i) W_i mostly appeared as class C in the training data; (ii) C_i exists in the class C gazetteer; (iii) W_{i-1} is a nationality; (iv) W_{i-1} is a quote; (v) W_{i-1} belongs to class C ; Finally, we would like to mention that the weights λ_i were estimated using the General Iterative Scaling (GIS) algorithm, which ensures convergence on the correct weights after a number of iterations. For this purpose we have used the YASMET⁹ software.

5 Experiments and Results

We have used the ANERcorp (see Section Three) to evaluate our system. The baseline model¹⁰ consists of assigning to a word W_i the class C_i which most frequently was assigned to W_i in the training corpus. However, as we have discussed in a previous paper [10], there is no available reference (neither a system nor a corpus) to compare our system with others. For this reason, we have used the demo version of the commercial system Siraj (Sakhr) and converted the obtained files to the IOB2 format to make possible the comparison with our system. We have used the CONLL 2002 evaluation software¹¹ which considers that a NE is correctly recognised only if: (i) all the constituent words of the NE are recognised; and (ii) the NE is correctly classified. Table 2 shows the baseline results. Table 3 illustrates the performance of the Siraj (Sakhr) system, whereas Tables 4 and 5 show the results obtained, respectively, by the first and the second version.

Table 2: Baseline results

Baseline	Precision	Recall	F-measure
Location	75.71%	76.97%	76.34
Misc	22.91%	34.67%	27.59
Organisation	52.80%	33.14%	40.72
Person	33.84%	14.76%	20.56
Overall	51.39%	37.51%	43.36

⁸ Basically unigrams and bigrams.

⁹ <http://www.fjoch.com/YASMET.html>

¹⁰ <http://cnts.ua.ac.be/conll2002/ner/bin/baseline>

¹¹ <http://bredt.uib.no/download/conllev.txt>

Table 3: Siraj (Sakhr) results

Siraj (Sakhr)	Precision	Recall	F-measure
Location	84.79%	67.91%	75.42
Misc	0.00%	0.00%	0.00
Organisation	0.00%	0.00%	0.00
Person	74.66%	55.84%	63.89
Overall	78.95%	46.69%	58.58

Table 4: ANERsys 1.0 results

ANERsys 1.0	Precision	Recall	F-measure
Location	82.17%	78.42%	80.25
Misc	61.54%	32.65%	42.67
Organisation	45.16%	31.04%	36.79
Person	54.21%	41.01%	46.69
Overall	63.21%	49.04%	55.23

Table 5: ANERsys 2.0 results

ANERsys 2.0	Precision	Recall	F-measure
Location	91.69%	82.23%	86.71
Misc	72.34%	55.74%	62.96
Organisation	47.95%	45.02%	46.43
Person	56.27%	48.56%	52.13
Overall	70.24%	62.08%	65.91

6 Discussion of Results

The results show clearly that ANERsys 2.0 performs more than 7 points (F-measure) better than the Siraj (Sakhr) system and significantly better than ANERsys 1.0. However, to make a deeper analysis of the results and have a clearer vision on ANERsys 2.0 we carried out some further experiments. Due to the two-steps approach adopted in the new version of our system we carried out three different tests. A first test to evaluate the performance of the first step of our new approach: i.e., the capacity of the system to delimit the NE's correctly (see Table 6). In order to evaluate the exact error rate of the second step, we used a corpus where the NE's delimitations were taken directly from the manually annotated corpus (see Table 7).

Table 6: Evaluation of the first step of the system

ANERsys 2.0	Precision	Recall	F-measure
B-NE	82.61%	72.10%	77.00
I-NE	91.27%	42.30%	57.81
Overall	84.27%	62.89%	72.03

Table 7: Evaluation of the second step of the system

ANERsys 2.0	Precision	Recall	F-measure
Location	93.22%	88.68%	90.90
Misc	94.67%	58.20%	72.08
Organisation	76.89%	65.27%	70.61
Person	75.10%	91.37%	82.44
Overall	83.22%	83.22%	83.22

The results illustrated above clearly that we need to improve the performance of the NE's delimitation process in order to enhance the performance of the complete system. The second step of the system gives an accuracy of 83.22: i.e., in case the first step was perfect the performance of our proposed system would be as good as the the best performance obtained in CONLL 2002 and 2003. Furthermore, it is also important to notice the our system performs better on the Person and Location classes which represent 69.4% of the NE's in our training corpus than the Miscellaneous and Organisation classes which represent only 30.6%. This shows that a greater training corpus will allow us to obtain a better performance.

7 Conclusions and Future Works

This paper presents a full description of the second version of ANERsys, an Arabic Named Entity Recognition system. In the first version of this system the ME approach proved to be convenient for the Arabic NER task. However, the system needed some improvement for a better multi-words named entities recognition. In order to tackle this problem, in the second version we have adopted a two-step approach where the first step concerns only the delimitation of the NE's using contextual and POS-tag information, whereas the second one, fully ME-based, completes the task by classifying the delimited NE's. Our new approach helped to raise more than 22 points above the baseline and 10 points over the last version. We carried out a comparison between our system and the available version of the Siraj (Sakhr) commercial system. The results showed that ANERsys 2.0 accuracy is almost 8 points greater than the other system on the F-measure. Furthermore, we present some further experiments which show that a better performance of our system depends on a better performance of the first step of the approach.

In ANERsys 2.0 the POS and BP tags were used to extract the NE boundaries determined by the POS tagger. In the next future, we plan to use the information provided by the POS tagger as additional features for the ME-based approach in order to assign a weight to the information obtained from the POS tagger before using them to determine the boundaries. We also plan to increase the size of the freely available ANERcorp and ANERgazet for a better quality of the system.

Acknowledgments

The research work of the first author was partially supported by MAEC - AECI. We would like to thank the PCI-AECI A/7067/06 and MCyT TIN2006-15265-C06-04 research projects for partially funding this work.

References

1. Abuleil, S. and Evens, M.: Names from Arab text for Question-Answering Systems. *Computers and the Humanities* Springer, 2002.
2. Maloney, J. and Niv, M.: TAGARAB, A Fast, Accurate Arabic Name Recognizer Using High-Precision Morphological Analysis. In *Proceedings of the Workshop on Computational Approaches to Semitic Languages*.
3. Cucerzan, S. and Yarowsky, D.: Language Independent NER using a Unified Model of Internal and Contextual Evidence. In: *Proceedings of CoNLL-2002, Taipei, Taiwan*, pp. 171-174.
4. Bender, O., Och, F.J. and Ney, H.: Maximum Entropy Models For Named Entity Recognition. In: *Proceedings of CoNLL-2003, Edmonton, Canada*, pp. 148-151.
5. Carreras, X., Márques, L. and Padró, L.: Named Entity Extraction using AdaBoost. In: *Proceedings of CoNLL-2002, Taipei, Taiwan* pp. 167-170.
6. Chieu, H.L., and Ng, H.T.: Named Entity Recognition with a Maximum Entropy Approach. In: *Proceedings of CoNLL-2003, Edmonton, Canada*, pp. 160-163.
7. Curran, J.R. and Clark, S.: Language Independent NER using a Maximum Entropy Tagger. In: *Proceedings of CoNLL-2003, Edmonton, Canada*, pp. 164-167.
8. Malouf, R.: Markov models for language-independent named entity recognition. *Proceedings of CoNLL-2002, Taipei, Taiwan*, pp. 187-190.
9. Florian, R., Ittycheriah, A., Jing, H., and Zhang, T.: Named Entity Recognition through Classifier Combination. In: *Proceedings of CoNLL-2003, Edmonton, Canada*, pp. 168-171.
10. Benajiba, Y., Rosso, P., Benedí Ruíz: ANERsys: An Arabic Named Entity Recognition System Based on Maximum Entropy. In: *Proceeding of CICLing-2007, Mexico. Lecture Notes in Computer Science 4394, Springer-Verlag*.
11. Diab, M., Hacıoglu, K. and Jurafsky, D.: Automatic Tagging of Arabic Text: From raw text to Base Phrase Chunks. In: *Proceedings of HLT-NAACL 2004*.