# Arabic Named Entity Recognition using Conditional Random Fields

**Yassine Benajiba and Paolo Rosso**

Natural Language Engineering Lab.
Departamento de Sistemas Informáticos y Computación
Universidad Politécnica de Valencia
Camino de Vera, s/n, 46022, Valencia (España)
{ybenajiba,prosso}@dsic.upv.es

**Abstract**

The Named Entity Recognition (NER) task consists in determining and classifying proper names within an open-domain text. This Natural Language Processing task proved to be harder for languages with a complex morphology such as the Arabic language. NER was also proved to help Natural Language Processing tasks such as Machine Translation, Information Retrieval and Question Answering to obtain a higher performance. In our previous works we have presented the first and the second version of ANERsys: an Arabic Named Entity Recognition system, whose performance we have succeeded to improve by more than 10 points, from the first to the second version, by adopting a different architecture and using additional information such as Part-Of-Speech tags and Base Phrase Chunks. In this paper, we present a further attempt to enhance the accuracy of ANERsys by changing the probabilistic model from Maximum Entropy to Conditional Random Fields which helped to improve the results significantly.

## 1. Introduction

The Named Entity Recognition (NER) task consists in determining and classifying proper names in an open-domain text. Many research works have been conducted to prove the predominant importance of NER to the other Natural Language Processing (NLP) tasks; some of these investigations are the following:

- In *Machine Translation* (MT), NEs require different techniques of translation than the rest of words of the text. Also, the post-editing step is more expensive when the errors of a MT system are mainly in NEs translation. For these reasons, (Babych and Hartley, 2003) have carried out a research study where they tag a text with a NER system as a pre-processing step of MT. The authors report that they have reached a higher accuracy with this new approach which helps the MT system to switch to a different translation technique when a Named Entity (NE) is detected.

- *Search results clustering*, is a sub-task of text clustering. It consists of organizing in groups the results returned by an IR system in order to make them easier to read for the user. In (Toda and Kataoka, 2005), the authors argue that they outperform the existing search results clustering techniques by including a NER system in their global system in order to give a special weight to the NEs in their clustering approach.

- *Information Retrieval* (IR), is a task which aims at retrieving the relevant document for the query formulated by the user in natural language: (Thompson and Dozier, 1997) report that 67.83%, 83.4% and 38.8% of the queries contained one or more Named Entities (NEs) according to Wall St. Journal, Los Angeles Times and Washington Post, respectively. Hence, an improvement of the retrieval of documents for queries which contain NEs would boost significantly the performance of the global IR system. In their research study, the authors have explored an approach which treats NEs and non-NEs differently. Their results show that the IR system precision outperforms the results obtained by a probabilistic retrieval engine on all the recall levels.

- *Question Answering* (QA), one of the most complicated NLP tasks because at satisfying the need of a special type of users which ask for an accurate answer to a specific question. Thus, a QA system does not stop at retreiving the relevant documents (like an IR system), it has also to answer but it has also to automatically extract the answer. In order to do so, a QA system has to perform several steps of processing both the question and the document-set where the system retrieves the answer (Benajiba et al., 2007). Many are the studies which show that the accuracy of a QA system relies significantly on the performance of the NER system included within, such as: (Ferrandez et al., 2007) which explore the accuracy of the global, both monolingual and cross-lingual, QA system for different NER systems. (Greenwood and Gaizauskas, 2007) use a NER system in order to improve the performance of an answer extraction module based on a pattern-matching approach. The authors use the NER system to capture the answers which are not possible to capture using only patterns. They report improving that the accuracy of answering the questions of type "When did X die" from 0% to 53%. (Mollá et al., 2006) also conducted a research study of the improvement obtained when the NER system tag-set corresponds exactly to the classes of NEs retrieved by the

QA system. The final results showed that up to 1.3% of improvement can be obtained in case both the NER system and global QA system aim at the same classes of NEs.

In order to use a standard definition of the NER task we have used the definition which was formulated in the in the shared task of the Conferences on Computational Natural Language Learning (CoNLL). In the sixth and the seventh editions of the Conference on Computational Natural Language Learning (CoNLL 2002[1] and CoNLL 2003[2]) the NER task was defined as to determine the proper names existing within an open domain text and classify them as one of the following four classes:

1. *Person*: named person or family;
2. *Location*: name of politically or geographically defined location;
3. *Organization*: named corporate, governmental, or other organizational entity; and
4. *Miscellaneous*: the rest of proper names (vehicles, weapons, etc.).

In the literature, very few research works were oriented especially to the NER task for Arabic texts (Abuleil, 2002; Maloney and Niv, 1998). Moreover, most of the effort were done for commercial purposes: Siraj[3] (by Sakhr), ClearTags[4] (by ClearForest), NetOwlExtractor[5] (by NetOwl) and InxightSmartDiscoveryEntityExtractor[6] (by Inxight). Unfortunately, no performance accuracy nor technical details have been provided and a comparative study of the systems is not possible. However, during the two editions of the CoNLL which we have previously mentioned, many research works addressed the language-independent NER task. A general study of these works showed that Maximum Entropy is an efficient approach for the task in question (Bender et al., 2003; Chieu and Ng, 2003; Curran and Clark, 2003; Cucerzan and Yarowsky, 1999; Malouf, 2003).

Recently, the Conditional Random Fields (CRF) model (Lafferty et al., 2001) proved to be very successful in many NLP tasks such as: shallow parsing (Sha and Pereira, 2003), morphological analysis (Kudo et al., 2004), information extraction (Pinto et al., 2003), biomedical NER (Settles, 2004), etc. Moreover, CRF proved a special success in the NER task for many languages of different levels of morphological complexity:

(i) *English* and *German*: (McCallum and Li, 2003) is one of the first attempts of using CRF for the NER task. The authors used the CoNLL 2003 corpus for evaluation and they report in their paper that an accuracy (F-measure) of 68.11 was reached for German, whereas 84.04 was obtained for English;

(ii) *Vietnamese*: in (Tran et al., 2007) a comparative study of Support Vector Machine (SVM) vs. CRF has been done and the results showed that using CRFs they have reached an accuracy of 86.48 vs. 87.75 using SVM. However, the authors report various experiments using different context window sizes for the SVM approach evaluation, whereas just one single result is reported for the CRF approach;

(iii) *Hindi*: 71.5 was reached for this language in (Li and McCallum, 2003) using CRF. However, the authors report that they have used a feature-induction technique because of their ignorance of the Hindi language peculiarities; and

(iv) *Chinese*: (Wu et al., 2006) reports in the paper that they have two different corpora for evaluation. For the first corpus, the best results were obtained when they used a combination of CRF and Maximum Entropy, whereas for the second corpus the best results were obtained for CRF. Moreover, the authors report that the worst results have been obtained when they have combined different CRF models.

To our knowledge, up to now there is no research study which has been carried out in order to prove the efficiency of the CRF model for NER in Arabic texts. Therefore, the idea behind the research work we present in this paper is to conduct experiments to investigate the performance of the CRF model for the Arabic NER task taking into consideration the peculiarities of the Arabic language and comparing the obtained results with our previous experiments which have been conducted using a Maximum Entropy approach.

The rest of this paper is structured as follows. In the second section of this paper we will give an overview of the Arabic language peculiarities. Section Three will describe our previous works related to the NER task. Section Four is dedicated to give a brief description of the CRF model. Details about the evaluation data we use in our experiments are given in Section Five. Finally, in the sixth section we present the results of our preliminary experiments with CRF and a comparison with our previous works results, whereas in the seventh section we draw some conclusions and discuss future works.

## 2. The Challenges of Arabic Named Entity Recognition

From a general viewpoint, the NER task can be considered as a composition of two sub-tasks:

1. *The detection of the existing NEs in a text* Which is a quite easy sub-task if we can use the capital letters as indicators to determine where the NEs start and where they end. However, this is only possible when the capital letters are supported in the target language, which is not the case for the Arabic language (Figure 1 shows the example of two words where only one of them is a NE and both of them start with the same character). The absence of capital letters in the Arabic language is the main obstacle to obtain high performance in NER (Benajiba et al., 2007)(Benajiba and Rosso, 2007).

2. *The classification of the NEs*

[1] http://www.cnts.ua.ac.be/conll2002/
[2] http://www.cnts.ua.ac.be/conll2003/
[3] http://siraj.sakhr.com/
[4] http://www.clearforest.com/index.asp
[5] http://www.netowl.com/products/extractor.html
[6] http://www.inxight.com/products/smartdiscovery/ee/index.php

(mouth) فم

(Valencia) فالنسيا

Figure 1: An example illustrating the absence of capital letters in Arabic

The Arabic language is a highly inflectional language, i.e., an Arabic word can be seen as the following composition:

*Word = prefix(es) + lemma + suffix(es)*

The *prefixes* can be articles, prepositions or conjunctions, whereas the *suffixes* are generally objects or personal/possessive anaphora. Both prefixes and suffixes are allowed to be combinations, and thus a word can have zero or more affixes. From a statistical viewpoint, this inflectional charactersitic of the Arabic language makes Arabic texts, compared to texts written in other languages which have a less complex morphology, more sparse and thus most of the Arabic NLP tasks are harder and more challenging. A full description of how thischaracterstic hardens each of the Arabic NLP goes beyond the scope of this paper. However, concerning the classification sub-task of NER, we can say that: the classification of NEs relies mainly on the word and the context in which it appeared in the text in order to decide the class it belongs to. Moreover, in case of an inflectional language, such as Arabic, both the words and the contexts may appear in different forms and thus a huge training corpus is required in order to obtain a high accuracy.

In order to reduce data sparseness in Arabic texts two solutions are possible:

(i) *Light stemming*: consists of omitting all the prefixes and suffixes which have been added to a lemma to obtain the needed meaning. This solution is convenient for tasks such as Information Retrieval and Question Answering because the prepositions, articles and conjunctions are considered as stop words and are not taken into consideration to decide whether a document is relevant for a query or not. An implementation of this solution was available on Kareem Darwish website[7] which has been unfortunately removed;

(ii) *Word segmentation*: consists of separating the different components of a word by a space character. Therefore, this solution is moreadequate for the NLP tasks which require to keep the different word morphemes such as Word Sense Disambiguation, NER, etc. A tool to perform Arabic word segmentation trained on Arabic Treebank, and obtaining an accuracy of *99.12* for this task, is available on Mona Diab website[8].

[7]http://www.glue.umd.edu/∼kareem/darwish

[8]http://www1.cs.columbia.edu/∼mdiab/

In our experiments we have adopted the second solution to reduce sparseness in our data and we draw the obtained results in the sixth section.

## 3. Our Previous Related Work

We have developed two versions of *ANERsys*, our Arabic NER system. Following we give a brief description of both versions of the system, whereas the results obtained with each of the systems will be given in the sixth section.

### 3.1. ANERsys 1.0: A Maximum Entropy Approach

As we have mentioned in the introduction of this paper, the Maximum Entropy approach has been very successful in the NER task. This approach is based on a exponential model which can be expressed as:

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

Z(x) is for normalization and may be expressed as:

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

Where *c* is the class, *x* is a context information and $f_i(x,c)$ is the i-th feature.

Maximum Entropy is a very convenient approach for the NER task thanks to its feature-based model. In this version of the system, our feature-set, which is fully *binary*, consisted of:

(i) *Wi*: The concerned word and its class;

(ii) $\{Wi\text{-}2, Wi\text{-}1\}$ *and* $\{Wi+1, Wi+2\}$: The bigrams coming before and after the word, which represent basically the context in which the word appears;

(iii) *Wi exists in a gazetteer*: The use of ANERgazet (see Section Five) as an external resource to enhance the system. The gazetteers were used in a binary way i.e., we have incorporated a binary feature which indicates whether *Wi* is an item of one of our gazetteers or not;

(iv) *Wi-1 is a nationality*: The NEs of class *person*, frequently come after the nationality of the person in question in newspapers articles.

### 3.2. ANERsys 2.0: A 2-step Approach

The error-analysis of ANERsys 1.0 results showed that the system had difficulties with multi-tokens NEs, i.e., it was harder to detect the Names Entities (NEs) than to classify them. Thus, in the second version of the system we have adopted a 2-step approach which is illustrated in Figure 2. The first step of the system is concerned mainly by detecting the start and the final tokens of each NE, whereas the second step takes care of classifying them (a full description of the system is given in (Benajiba and Rosso, 2007))

## 4. Conditional Random Fields

CRFs (Lafferty et al., 2001) is a probabilistic framework to segment and label sequence data. It is based on undirected graphical models where the nodes represent the label sequence **y** corresponding to the sequence **x**. CRF model aims at finding the label **y** which maximizes the conditional probability **p(y|x)** for a sequence **x**. The CRF model is

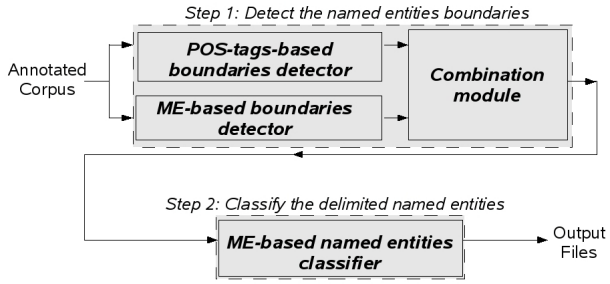


Figure 2: Generic architecture of ANERsys 2.0

a feature-based model where features have binary values such as:

*$f_k(y_{t-1}, y_t, x)$:=1 for x='Darfur' and $y_t$='B-LOC', and 0 otherwise.*

The CRF model is considered a generalization of Maximum Entropy and Hidden Markov Models (HMM) and can be expressed as following:

$$p(y|x) = \frac{1}{Z(x)} * exp(\sum_t \sum_k \lambda_k . f_k(y_{t-1}, y_t, x)) \quad (3)$$

where $\lambda_i$ represent the weights assigned to the different features in the training phase and **Z(x)** is a normalization factor which can be expressed as:

$$Z(x) = \sum_{y \epsilon Y} exp(\sum_t \sum_k \lambda_k . f_k(y_{t-1}, y_t, x)) \quad (4)$$

## 5. Evaluation Data

We have used ANERcorp in order to train and test the CRF model. ANERcorp is composed of a training corpus and a test corpus annotated especially for the NER task. We have chosen the tokens of ANERcorp from both news wire and other web resources (more details about ANERcorp are given in (Benajiba et al., 2007)) and we have manually annotated them ourselves. Each token of ANERcorp is tagged as belonging to one of the following classes:

- B-PERS: The Beginning of the name of a PERSon.
- I-PERS: The continuation (Inside) of the name of a PERSon.
- B-LOC: The Beginning of the name of a LOCation.
- I-LOC: The Inside of the name of a LOCation.
- B-ORG: The Beginning of the name of an ORGanization.
- I-ORG: The Inside of the name of an ORGanization.
- B-MISC: The Beginning of the name of an entity which does not belong to any of the previous classes (MISCellaneous).
- I-MISC: The Inside of the name of an entity which does not belong to any of the previous classes.
- O: The word is not a named entity (Other).

ANERcorp contains more than 150,000 tokens (11% of the tokens are part of a NE) and they are freely downloadable from our website[9]. The ANERcorp has been used in our earlier work (Benajiba et al., 2007) (Benajiba and Rosso, 2007) in order to evaluate the two versions of ANERsys which we have described before (see Section Three).

## 6. Experiments and Results

### 6.1. Corpus, Baseline, Measure

We have used the ANERcorp (see Section Five) to evaluate our system. The baseline model[10] consists of assigning to a word $w_i$ the class $C_i$ which most frequently was assigned to $w_i$ in the training corpus. The words which were unseen during the training phase are assigned the class *O*. We have used the $F_{\beta=1}$-measure for evaluation:

$$F_{\beta=1} = \frac{(\beta^2 + 1) * precision * recall}{\beta^2 * (precision + recall)} \quad (5)$$

Where *precision* is the percentage of NEs found by the system and which are correct. It can be expressed as:

$$precision = \frac{Num(correct\ NEs\ found)}{Num(NEs\ found)} \quad (6)$$

and *recall* is the percentage of NEs existing in the corpus and which were found by the system. It can be expressed as:

$$recall = \frac{Num(NEs\ found)}{Total\ number\ of\ NEs} \quad (7)$$

### 6.2. Feature-set

We have kept the same feature-set used in our previous systems (see Section Three) in order to be able to compare the performance of the Maximum Entropy (ME) and the CRF performance.

**POS-tag and BPC** : The Part-Of-Speech tagging is the task of assiging to each word its linguistic category. Base Phrase Chunks (BPC) are atomic parts of a sentence (beyond words). In CoNLL 2003, the POS-tags, together with the BPC, formed part of the corpora which were provided to the participants (see Figure 3). The point of using POS-tags and BPS relies mainly on that BPC might determine the beginning and the end of a NE and thus help the classifier to capture the boundaries of the NEs. Additionally, using the POS-tags is also helpful thanks to the "*NNP*" tag which marks a word a NE. However, in the proceedings of the conference there were no studies reporting the impact of each of these features individually.

[9] http://www.dsic.upv.es/∼ybenajiba

[10] http://cnts.ua.ac.be/conll2002/ner/bin/baseline

| | | | |
|---|---|---|---|
| U.N. | NNP | I-NP | I-ORG |
| official | NN | I-NP | O |
| Ekeus | NNP | I-NP | I-PER |
| heads | VBZ | I-VP | O |
| for | IN | I-PP | O |
| Baghdad | NNP | I-NP | I-LOC |
| . | . | O | O |

Figure 3: An extract of the CoNLL 2003 English corpus

**External Resources (*GAZ*)** : In order to measure the impact of using external resources in the NER task we have used ANERgazet (also available on our website) which consists of three different gazetteers, all built manually using web resources:
(i) *Location Gazetteer*: this gazetteer consists of 1,950 names of continents, countries, cities, rivers and mountains found in the Arabic version of wikipedia[11];
(ii) *Person Gazetteer*: this was originally a list of 1,920 complete names of people found in wikipedia and other websites. After splitting the names into first names and last names and omitting the repeated names, the list contains finally 2,309 names;
(iii) *Organizations Gazetteer*: the last gazetteer consists of a list of 262 names of companies, football teams and other organizations.

$W_{i-1}$ **is a Nationality (*NAT*)** : Frequently, NEs of the class "Person" comes after mentioning the nationality of the person (especially in newspaper articles). For instance, *the Iranian Presiden Mahmoud declared ....*

### 6.3. Results

**Baseline and Previous Results** Table 1 shows the baseline results. Tables 2 and 3 show the results obtained, respectively, by the first and the second version of ANERsys. Using a ME approach (ANERsys 1.0) has helped to obtain an F-measure which is almost 12 points above the baseline (55.23). Moreover, when we have used a 2-step approach and adopted different techniques for detecting and classifying the NEs, we have significantly raised the recall of our system from 49.04% to 62.08%, and hence the performance of the system was enhanced and has reached an F-measure of 65.91.

Table 1: 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** |

[11] http://ar.wikipedia.org

Table 2: 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 3: 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** |

**Impact of Tokenization** In our previous works, the error-rate induced by the complex morphology of the Arabic language was not taken into consideration. This error-rate is mainly due to the bad training which is a direct consequence of the sparseness of data caused by the agglutinative morphology. In this paper, we have conducted experiments before and after the tokenizing the data. In Table 4 we present the results obtained with raw text, whereas the results obtained after the tokenization, are presented in Table 5, using CRF (we have used CRF++[12]).

Table 4: CRF results using non-tokenized data

| CRF Raw | Precision | Recall | F-measure |
|---|---|---|---|
| Location | 95.09% | 70.02% | 80.65 |
| Misc | 78.31% | 50.39% | 61.32 |
| Organisation | 85.27% | 46.51% | 60.19 |
| Person | 80.18% | 36.73% | 50.38 |
| Overall | **89.20**% | **54.63**% | **67.76** |

Table 5: CRF results using tokenized data

| CRF Tok. | Precision | Recall | F-measure |
|---|---|---|---|
| Location | 95.38% | 76.14% | 84.68 |
| Misc | 79.49% | 47.33% | 59.33 |
| Organisation | 86.28% | 48.28% | 61.92 |
| Person | 84.87% | 38.18% | 52.67 |
| Overall | **90.82**% | **57.83**% | **70.67** |

**Features** The rest of the tables show the results obtained using each of the features individually and then combining all of them. Table 6, 7, 8 and 9 show the impact of the POS-tag, BPC, GAZ and NAT, respectively.

[12] http://crfpp.sourceforge.net/

Table 6: Results obtained using the POS-tag feature

| POS | Precision | Recall | F-measure |
|---|---|---|---|
| Location | 89.88% | 86.49% | 88.15 |
| Misc | 77.91% | 51.15% | 61.75 |
| Organisation | 83.02% | 53.33% | 64.94 |
| Person | 79.29% | 65.42% | 71.69 |
| Overall | **85.28**% | **71.82**% | **77.97** |

Table 7: Results obtained using the BPC feature

| BPC | Precision | Recall | F-measure |
|---|---|---|---|
| Location | 95.97% | 77.28% | 85.62 |
| Misc | 80.25% | 49.62% | 61.32 |
| Organisation | 85.87% | 49.09% | 62.47 |
| Person | 86.39% | 41.52% | 56.09 |
| Overall | **91.35**% | **59.62**% | **72.15** |

Table 8: Results obtained using the GAZ feature

| GAZ | Precision | Recall | F-measure |
|---|---|---|---|
| Location | 94.36% | 79.21% | 86.12 |
| Misc | 81.58% | 47.33% | 59.90 |
| Organisation | 85.66% | 48.28% | 61.76 |
| Person | 84.94% | 43.66% | 57.67 |
| Overall | **90.22**% | **60.85**% | **72.68** |

Table 9: Results obtained using the NAT feature

| NAT | Precision | Recall | F-measure |
|---|---|---|---|
| Location | 95.60% | 76.32% | 84.88 |
| Misc | 79.75% | 48.09% | 60.00 |
| Organisation | 84.86% | 48.69% | 61.87 |
| Person | 85.80% | 40.32% | 54.86 |
| Overall | **90.83**% | **58.66**% | **71.29** |

Table 10: Results obtained combining "all" the features

| ALL | Precision | Recall | F-measure |
|---|---|---|---|
| Location | 93.03% | 86.67% | 89.74 |
| Misc | 71.00% | 54.20% | 61.47 |
| Organisation | 84.23% | 53.94% | 65.76 |
| Person | 80.41% | 67.42% | 73.35 |
| Overall | **86.90**% | **72.77**% | **79.21** |

## 7. Results Discussion and Error Analysis

**By Features** : When each feature was used individually, the POS-tag (Table 6) feature showed the best improvement in F-measure (more than 7 points). The contribution of the POS-tag feature was mainly on the recall (amost 14 points), whereas for the precision it has caused a significant decrease (more than 5 points). The only feature which showed to help increasing the precision is the BPC feature (Table 7). However, the improvement in both precision and recall was very light. Using external resources has only helped to increase 3 points in recall (Table 8), whereas for the NAT feature, it has contributed with an improvement of 0.62 points (Table 9).

**By Classes** : The CRF model has benefited from all the features for all the classes. However, the results tables show that all the classes have benefited more from the POS-tag feature than the other features on the recall and F-measure levels. On the other hand, the "Location", "Organization" and "Person" classes show that they gain more in precision with the BPC feature, whereas the "Miscellaneous" class improves more in precision with the GAZ feature. The major difference between the "Miscellaneous" class and the other classes is that the contexts in which its potential subclasses (weapons, currencies, vehicles, etc.) might appear are very different. On the other hand, the NEs which belong to the other classes are more precisely defined and even though they have sub-classes (Person: president, actor, etc. Location: country, city, street, etc. Organization: research center, soccer team, fashion label, etc.) they tend to appear in the same context. For this reason, the "Miscellaneous" class benefits more from using external resources than using other features.

**Combination of the Features** : When all the features were combined (Table 10), the obtained recall (72.77%) was almost one point above the best recall obtained by a single feature (71.82%, see Table 6), whereas the precision was (86.90%) almost 4 points below the best precision obtained when the BPC feature was used individually (91.35%). However, on the F-measure level, Table 10 shows that the performance is almost 2 points above using only the POS-tag feature. That is, when a CRF model is user with independent features of different types in the NER task, it succeeds to combine these features and obtain results which outperform the ones obtained when these features are used individually.

## 8. Conclusions and Further Work

In this paper we present our preliminary experiments which aim at improving ANERsys, our NER system for Arabic text, by using the CRF model.

The results showed that with the CRF model we can obtain a performance almost two points higher with respect to the second version of ANERsys which relies on a 2-step approach and partially on a Maximum Entropy model. Due to the complex morphology of the Arabic language, we have performed a tokenization on our data which helped to gain almost three points. Thereafter, we have performed experiments using four different gazetteers individually and combining them. The results showed that we have obtained more improvement in recall than in precision. Moreover, some classes ("Miscellaneous") showed that they benefit more from using external resources than morphological (POS-tag) feature. When all the features were combined, the CRF models showed that it outperfoms other probabilistic model in the ability to capture arbitrary, overlapping features (Kristjansson et al., 2004). The overall F-measure

was enhanced more than one point above the best result obtained using only one feature (POS-tag), almost 9 points above the results obtained when no features were added and almost 14 points above the results obtained with the second version of our Arabic NER system (65.21). All the features that we have used in our experiments are language-independent which will allow many NLP researchers to benefit from our research work for othe languages.
In the next future we plan to increase the size of ANERcorp in order to obtain a higher performance of the system. We also plan to carry out experiments using different feature-sets, and explore the possibility of designing a feature-set for each class. Furthermore, we plan to conduct a comparative study between many probabilistic models (SVM, HMM, Maximum Entropy, CRF, etc.) and also experiments using a combination of different models.

**Acknowledgments**

The research work of the first author was partially supported by MAEC - AECI. We would like to thank the PCI-AECI A/010317/07 and MCyT TIN2006-15265-C06-04 research projects for partially funding this work.

## 9. References

Abuleil S. 2002. *Extracting Names from Arabic text for Question-Answering Systems. Computers and the Humanities.*

Babych B. and Hartley A. 2003. *Improving Machine Translation Quality with Automatic Named Entity Recognition*. In *Proc. of EACL-EAMT*. Budapest.

Benajiba Y., Rosso P., Benedí J.M. 2007. *ANERsys: An Arabic Named Entity Recognition system based on Maximum Entropy.*. In *In: Proc. 8th Int. Conf. on Comput. Linguistics and Intelligent Text Processing, CICLing-2007, Springer-Verlag, LNCS(4394), pp. 143-153.*.

Benajiba Y., Rosso P. 2007. *ANERsys 2.0 : Conquering the NER task for the Arabic language by combining the Maximum Entropy with POS-tag information.*. In *In: Proc. Workshop on Natural Language-Independent Engineering, 3rd Indian Int. Conf. on Artificial Intelligence, IICAI-2007, Pune, India, December 17-19*.

Benajiba Y., Rosso P. and Lyhyaoui A. 2007, *Implementation of the ArabiQA Question Answering System's Components*, *In Proc. of ICTIS-2007*,

Bender O., Och F., and Ney H.. 2003. *Maximum Entropy Models For Named Entity Recognition*. In *Proceedings of CoNLL-2003*. Edmonton, Canada.

Chieu H. and Ng H. 2003. *Named Entity Recognition with a Maximum Entropy Approach*. In *Proceedings of CoNLL-2003*. Edmonton, Canada.

Cucerzan S. and Yarowsky D. 1999. *Language Independent Named Entity Recognition Combining Morphological and Contextual Evidence*. In *Proceedings, 1999 Joint SIGDAT Conference on Empirical Methods in NLP and Very Large Corpora*, pp. 90–99.

Curran J. R. and Clark S. 2003. *Language Independent NER using a Maximum Entropy Tagger*. In *Proceedings of CoNLL-2003*. Edmonton, Canada.

Ferrndez S., Ferrndez O., Ferrndez A. and Muoz R., 2007. *The Importance of Named Entities in Cross-Lingual Question Answering*, *In Proc. of Recent Advances in Natural Language Processing, RANLP-2007*. Borovets, Bulgaria.

Greenwood M. and Gaizauskas R. 2007. *Using a Named Entity Tagger to Generalise Surface Matching Text Patterns for Question Answering*, *In Proceedings of the Workshop on Natural Language Processing for Question Answering (EACL03)*,

Kristjansson T., Culotta A., Viola P., and McCallum A. 2004. *Interactive Information Extraction with Constrained Conditional Random Fields*. In *Proceedings of AAAI-2004*.

Kudo T., Yamamoto K., and Matsumoto Y. 2004. *Applying Conditional Random Fields to Japanese Morphological Analysis*. In *Proceedings of EMNLP*, 2004.

Lafferty J., McCallum A., and Pereira F. 2001. *Conditional random fields: Probabilistic models for segmenting and labeling sequence data*. In *Proceedings of the Eighteenth International Conference on Machine Learning*. 2001.

Li W. and McCallum A. 2003. *Rapid Development of Hindi Named Entity Recognition using Conditional Random Fields and Feature Induction*. In *Special issue of ACM Transactions on Asian Language Information Processing: Rapid Development of Language Capabilities: The Surprise Languages*.

Maloney J. and Niv M. 1998. *TAGARAB, A Fast, Accurate Arabic Name Recognizer Using High-Precision Morphological Analysis*. In *Proceedings of the Workshop on Computational Approaches to Semitic Languages*.

Malouf R. 2003. *Markov Models for Language-Independent Named Entity Recognition*. In *Proceedings of CoNLL-2003*. Edmonton, Canada.

McCallum A. and Li W. 2003. *Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons*. In *Proceedings of the Seventh Conference on Natural Language Learning (CoNLL)*.

Mollá D., van Zaanen M. and Smith D. 2006. *Named Entity Recognition for Question Answering*, *Proc. of the Australasian Language Technology Workshop Sancta Sophia College*

Pinto D., McCallum A., Wei X., and Croft W. B. 2003. *Table Extraction Using Conditional Random Fields*. In *Proceedings of the 26th ACM SIGIR.*, 2003.

Settles B. 2004. *Biomedical Named Entity Recognition Using Conditional Random Fields and Rich Feature Sets*. In *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications (NLPBA)*.

Sha F. and Pereira F. 2003. *Shallow parsing with conditional random fields*. In *Proceedings of HLT-NAACL*.

Strótgen R., Mandl T. and Schneider R. 2005. *A Fast Forward Approach to Cross-Lingual Question Answering for English and German*. In *Proceedings of the Workshop of Cross-Language Evaluation Forum (CLEF)*. 2005.

Thompson P. and Dozier C., 1997. *Name Searching and Information Retrieval*, *In Proc. of Second Conference on Empirical Methods in Natural Language Processing*,

Toda H. and Kataoka R., 2005. *A Search Result Clustering Method using Informatively Named Entities*, *In Proc. of the 7th annual ACM international workshop on Web information and data management.*,

Tran Q. T., Pham T. X. T., Ngo Q. H., Dinh D., and Collier N. 2007. *Named Entity Recognition in Vietnamese documents*. *Progress in Informatics Journal*. 2007.

Wu C-W., Jan S-Y., Tsai R. T-H., and Hsu W-L. 2006. *On Using Ensemble Methods for Chinese Named Entity Recognition*. *Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing*. 2006.