Using Language Independent and Language Specific Features to Enhance Arabic Named Entity Recognition

Yassine Benajiba[2], Mona Diab[2], and Paolo Rosso[1]
[1]Natural Language Engineering Laboratory, ELiRF, Universidad Politécnica Valencia, Spain
[2]Center of Computational Learning Systems, Columbia University, USA

Abstract: *The Named entity recognition task has been garnering significant attention as it has been shown to help improve the performance of many natural language processing applications. More recently, we are starting to see a surge in developing named entity recognition systems for languages other than English. With the relative abundance of resources for the Arabic language and a certain degree of maturation in the state of the art for processing Arabic, it is natural to see interest in developing NER systems for the language. In this paper, we investigate the impact of using different sets of features that are both language independent and language specific in a discriminative machine learning framework, namely, Support Vector Machines. We explore lexical, contextual and morphological features and nine data-sets of different genres and annotations. We systematically measure the impact of the different features in isolation and combined. We achieve the highest performance using a combination of all features, F1=82.71. Essentially combining language independent features with language specific ones yields the best performance on all the genres of text we investigate. However, on a class level, we observe that the different classes of named entities benefit differently from the morphological features employed.*

Keywords: *Arabic natural language processing, classification, information extraction, named entity recognition.*

Received December 18, 2008; accepted June 21, 2009

1. Introduction

The Named Entity Recognition (NER) task is one of the most important subtasks in information extraction. It is defined as the identification and classification of Named Entities (NE's) within an open-domain text. We find significant research that covers a large variety of techniques used for efficient NER systems [19, 16, 21]. Thanks to standard evaluation test beds such as the Automatic Content Extraction (ACE), the task of NER has garnered significant attention within the natural language processing community. ACE has facilitated evaluation for different languages by creating standardized test sets and evaluation metrics.

NER systems are typically enabling subtasks within large Natural Language Processing (NLP) systems. The quality of the NER system has a direct impact on the quality of the overall NLP system. Evidence abound in the literature in areas such as Question Answering (QA) task, the majority of the considered questions at the TREC and CLEF competitions expect a NE or a list of NEs as answers [6]. In clustering search results, the use of a NER system before comparing the documents contents proved to be very useful [23]. In machine translation, [2] has shown that NER pre-processing improves the quality of the translation output.

In this paper, we address the problem of NER for Arabic. The NER task in Arabic is relatively different from performing the task in English due to the inherent characteristic linguistic differences of Arabic, most notably, the lack of a direct signal such as capitalization in Arabic orthography to mark a named entity.

We adopt a discriminative approach to the NER problem. We use Support Vector Machines (SVMs) [24]. We comprehensively investigate many sets of features: contextual, lexical, morphological and shallow syntactic features. We explore the features in isolation as well as in combination with each other. We experiment with two sets of data, the standard ACE data and a manually created data set UPV-corpus. Our best system that combines all the features yields an overall F1 score of 82.71.

The paper is structured as follows. Section 2 gives a general overview of the state-of-the-art NER approaches with a particular emphasis on Arabic NER; section 3 describes relevant characteristics of the Arabic language illustrating the challenges posed to NER; in section 4, we discuss the details of our approach including the different tag sets and feature-sets; section 5 describes the experiments and shows the results obtained; finally, we discuss the results and some of our insights in section 6.

2. Related Work

There are several significant research efforts in NER. In the Conference on Natural Language Learning (CoNLL) 2002 and 2003 NER evaluation tasks, respectively, the most successful language independent approaches to NER are systems that employ Maximum Entropy (ME) techniques in a supervised setup [5, 7, 8]. Malouf, in [19], investigated the difference in performance between Hidden Markov Models (HMM) and ME. He shows the superiority of the ME approach to the problem of NER for English.

NER for other languages, such as Hindi [16] and Chinese [25], have explored Conditional Random Fields (CRF) successfully. However, [21] show that using a SVMs approach outperforms (F1=87.75) using CRFs (86.48) on the NER task in Vietnamese.

With current surge in resources for Arabic making their way in the NLP community, we are starting to see systems being developed for the processing of the Arabic language. Earlier systems relied on rule based methods to solve the problem. For instance, [18] combines a morphological analyzer and a pattern matching module. Whereas [1] developed a system that was entirely based on hand-crafted rules and triggers. It is difficult to compare the performance of these different systems as they did not use standard test sets, tag sets or corpora.

More recently, we have shown in our work [3] that using a basic ME approach to Arabic NER yields an F1-measure of 55.23. We followed up with further work in [4], we report results reaching F1=65.91 by adopting a two stage classification approach to the NER problem. This approach divides the NER task into two subtasks: a NE boundary detection task; and an NE classification task. However, in that work, we did not exploit any of the characteristic features of the Arabic language and we did not evaluate against standard data sets.

In work by Farber *et al*. [12], they use a perceptron model and evaluate their results against standard ACE data sets for Arabic. They report no gain from using morphological features in their experiments, however, they do show improvement in the results (*F* measure of 71.5 on ACE 2005 data set) if they use the output of the morphological disambiguator to choose the correct Arabic analysis which contains the English translation. If the English translation is capitalized, then this is a cue that the word in Arabic is a NE.

Similarly, in work by [26], the authors investigate the Arabic mention detection problem. A mention can be a named, e.g., Ohio, a nominal, e.g., Prime Minister, or a pronominal, e.g., he reference to an entity. First the data is pre-processed applying morphological stemming. For classification, the authors implemented a Maximum Entropy Markov Model approach using lexical, syntactic and gazetteer features. The authors evaluate their system's performance against the ACE 2004 data. Their system yields an overall *F*-measure of 69. However, the result is not broken down by the different types of mention. Therefore it is hard to tell what the performance on the NE alone was.

3. Arabic in the Context of the Named Entity Recognition Task

3.1. The Arabic Language

Arabic is a Semitic language. It is known for its templatic morphology where words are made up of roots and affixes. Clitics agglutinate to words. For instance, the surface word وبحسناتهم *wbHsnAthm* `and by their virtues[fem.]', can be split into the conjunction *w* `and', preposition *b* `by', the stem *HsnAt* `virtues [fem.]', and possessive pronoun *hm* `their'.

3.2. Challenges

There exist two major challenges posed by focusing on Arabic NER:

- Absence of capital letters in the orthography: English like many other Latin script based languages has a specific signal in the orthography, namely capitalization of the initial letter, indicating that a word or sequence of words is a named entity. Arabic has no such special signal rendering the detection of NEs more challenging. In Arabic, there is no such orthographic marker for NEs.
- The Arabic language is highly inflectional: as we mentioned earlier, Arabic language uses an agglutinative strategy to form surface tokens. As seen in the example above, a surface Arabic word maybe translated as a phrase in English. Consequently, the Arabic data in its raw surface form (a statistical viewpoint) is much more sparse which, in a learning setting, decreases the efficiency of the training significantly. In order to tackle this problem, it is needed to perform a segmentation of the clitics of each word (tokenization) as a pre-processing step.

It helps particularly for the NER task to overcome two major difficulties:

- Make the NEs appear always in the same form (which lowers the number of unseen NEs).
- Reduce the number of surface forms of the contexts in which the NEs appear.

4. Approach Using a Large Range of Features

4.1. An SVM-based Approach

SVMs [24] goal is to find the hyperplane *H* which separates the elements of two classes (-1 and +1) with a maximum margin between *H* and the closest data

point to it. *H* is computed in the training phase during the observation of a set of elements of each class. Those training data are represented with their vector of features f_i and their class c. In the test phase, SVMs assigns the class +1 to an element if it is located on the 'plus' side of *H* and -1 if it is on the 'minus' one. The position of an element is computed using the following formula:

$$g(x)=\Sigma_i^N w_i.k(x,sv_i) + b \quad (1)$$

where sv_i are the support vectors which are the nearest elements to *H*, *N* is the number of support vectors, $k(x; sv_i)$ are the so-called kernels which map the features vectors into a another space, w_i are the weights assigned to the different features representing an element and *b* is a constant which is also determined in the training phase.

SVMs are robust to noise in the data and they have powerful generalization ability especially in the presence of a large number of features. Moreover, SVMs have been used successfully in many NLP areas of research in general [9, 10, 13, and 15], and for the NER task in particular [21, 20].

In this paper we employ a sequence model over SVMs, Yamcha that converts the NER task to a chunking task using the Inside Outside Beginning (IOB) tagging scheme. Table 1 illustrates the tagging scheme in the IOB format.

Table 1. An example of an IOB2 annotated corpus.

Arabic	Buckwalter	English Translation	Tag
وفي	Wfy	and in	O
بيروت	Byrwt	Beirut	B-LOC
,	,	,	O
وصف	wSf	described	O
فؤاد	F&Ad	Fouad	B-PER
السنيورة	Alsnywrp	Siniora	I-PER
رئيس	R}ys	president	O
الوزراء	AlwzrA'	the ministers	O
اللبناني	AllbnAny	Lebanese	O

4.2. Arabic NER Task Tag Sets

Since inherently the NER task is an enabling technology, we acknowledge that different NLP applications may require different tag sets, however, here we address the NER task as an end system in itself. There exist three standard NER tag sets in the literature:

- The Message Understanding Conference (MUC-6): the NER task consisted of three subtasks: ENAMEX for proper nouns, NUMEX for numerical expressions and TIMEX for temporal expressions. The ENAMEX subtask is defined as the identification of the NE's and their classification into: person, e.g., Albert Einstein, location, e.g., paris, or organization, e.g., Google Co..
- CoNLL: in the language-independent NER shared task held in the CoNLL2002 and CoNLL2003 the tag-set comprised four classes: person, location, organization same as MUC-6 definitions and Miscellaneous, e.g., empire state building.
- ACE: the ACE 2003 data defines four different classes: person, Geographical and Political Entities (GPE), organization and facility. Moreover, in ACE 2004 and 2005 two classes were added to the tag-set: Vehicles, e.g., Rotterdam ship and weapons, e.g., Kalashnikov.

We note that the three data sets include person, location in the ACE set this corresponds to the more specified geographical and political entity and Organization. ACE adds facility, vehicles and weapons, while CoNLL has a Miscellaneous category. Even though some of these sets use the same tags, the definitions and the scope of what constitutes a NE differ from one gold standard set to the other.

4.3. Features

The most challenging aspect of any machine learning approach to NLP problems is deciding on an optimal feature set. In this work, we investigate a large space of features characterized as follows.

- ConteXTual (CXT): this is an automatically generated feature that accounts for the different contexts in which NEs appear in the training data. The context is defined as a window of +/- n tokens from the NE of interest.
- LEXical (LEX): this feature defines the lexical orthographic nature of the tokens in the text. These features include: special markers for tokens that include digits or punctuation as is the case with abbreviations which contain periods; number of characters in a token; a window of different sized n-gram sequence characters for the tokens.
- GAZetteers (GAZ): these include hand-crafted dictionaries/gazetteers listing predefined NEs. We use three gazetteers for people, locations and organization names. We semi-automatically enriched the location gazetteer using the Arabic wikipedia as well as other web sources. This enrichment consisted of:
- Taking the page labeled 'countries of the world (دول العالم) as a starting point to crawl into Wikipedia and retrieve location names.
- We automatically filter the data removing stop words.
- Finally, we manually filter the resulting set ensuring its good quality as a source of location names.
- MORPHological features (MORPH): this feature set is based on exploiting the rich characteristic morphological features of Arabic. We relied on a system for Morphological Analysis and Disambiguation for Arabic (MADA) to extract relevant morphological information [13]. MADA yields an accuracy of 95% on morphological

disambiguation. Arabic morphology is complex exhibiting both derivational and inflectional morphology. MADA disambiguates words along 14 different morphological dimensions. MADA typically operates on untokenized texts (surface words as they naturally occur), hence several of the features indicate whether there are clitics of different types. We use MADA for the preprocessing step of clitic tokenization.

The features produced by MADA that are of most relevance to us in the NER task are the morphological features that affect nominals such as case, number, gender, person, and definiteness. Proper names in general do not inflect and they rarely exhibit case information, therefore the lack of these morphological features is an indicative signal. We use the MADA features in two different ways:

- Without making any changes ($MORPH_{raw}$).
- Grouping some of the MADA features together to emphasize their discerning power for NE detection ($MORPH_{mod}$).

Part Of Speech (POS) tags and Base Phrase Chunks (BPC): to derive part of speech tags POS and base phrase chunks BPC we employ the AMIRA-1.0 system described in [11]. Like the MADA system, AMIRA-1.0 is an SVMs-based set of tools. The POS tagger performs at 96.2% and the BPC system performs at 95.41%. It is worth noting here that the MADA system produces POS tags however it does not produce BPC, hence the need for a system such as AMIRA-1.0. We use the reduced LDC BIES POS tag set of 25 tags created for the Arabic Treebank [17].

- NATionality (NAT): this feature is both a contextual and a lexical feature. We mark nationalities in the input text. Such information is useful in detecting NEs as they are used as precursors to recognizing NE. For instance, we mark the abundance of the following type of structure 'وصرح الرئيس الايراني محمود...' wSrH Alr{ys AlAyrAny mHmwd ...' corresponding to `and the iranian president Mahmoud declared ...', where a NE is preceded by a nationality.
- Corresponding English CAPitalization (CAP): MADA provides the English translation for the words it morphologically disambiguates as a side effect of running the morphological disambiguation. In the process it taps into an underlying lexicon that provides bilingual information. Inspired, the insight is that if the translation begins with a capital letter, then it is most probably a NE. This feature is used for both investigating its usefulness to enhance NER performance in our experimental setting as well as investigate the impact of the absence of such a signal from the orthography on NER results.

5. Experiments and Results

5.1. Data

We use the ACE 2003, 2004 and 2005 corpora and an enhanced version of the corpus used in our previous work [3], UPV-corpus. Table 2 describes the characteristics of the different corpora: the training data size ($Size_{train}$), the test size ($Size_{test}$), the ratio of the NE tokens to the total number of tokens ($Ratio_{NE}$), the sum of the number of NEs in the training and the test corpora (Num_{NE}), and the average of the number of tokens in a NE (Avg_{length}).

UPV-Corpus: this corpus comprises text collected from different newswire web sources. The texts are manually annotated. Several rounds of reviews are performed to ensure the consistency of the data. The tag set used is the same as the CoNLL tag set described in section 4.2.

ACE data: the ACE data is annotated for many tasks: Entity Detection and Tracking (EDT), Relation Detection and Recognition (RDR), Event Detection and Recognition (EDR). In 2003, there were two genres of data: Broadcast News (BN) and NewsWire (NW). An additional genre, Arabic TreeBank (ATB), was added in 2004. In 2005, the ATB genre was replaced by WebLogs (WL). A main difference between the annotations of the UPV-corpus and ACE annotations is that in the latter, nationalities (e.g. Spanish, French, *etc.*) are tagged as GPE whereas in the former they are not considered NEs.

Table 2. Characteristics of UPV-corpus and ACE 2003, 2004 and 2005 data.

	UPV-corpus	ACE 2003		ACE 2004			ACE 2005		
	NW	BN	NW	BN	NW	ATB	BN	NW	WL
$Size_{train}$	144.48k	16.34k	29.44k	50.44k	51.74k	21.27k	22.3k	43.85k	18k
$Size_{test}$	30.28k	2.51k	7k	13.32k	13.4k	5.25k	5k	12.3k	3.2k
$Ratio_{NE}$	10.94%	11.49%	11.81%	11.49%	11.81%	12.58%	19%	15.41%	6.56%
Num_{NE}	12989	2100	3405	4609	4839	2072	3553	5697	968
Avg_{length}	1.48	1.32	1.43	1.59	1.59	1.60	1.46	1.52	1.43

In order to carry out our experiments correctly with the ACE data, we remove all annotations which are not oriented to the EDR task. Hence, all the listed characteristics for this corpus pertain to the portions of the data that are relevant to NER only.

5.2. Experimental Set Up

5.2.1. Metrics

We use the CoNLL evaluation standard metrics of precision, recall and *F1*-measure [22]. The CoNLL evaluation metric is an aggressive metric that does not assign partial credit. An NE has to be identified as a whole and correctly classified in order to gain credit.

5.2.2. Experiments

We have three sets of experiments in this paper: a baseline, a parameter setting set of experiments, and then feature engineering experiments.

- Baseline: we use the CoNLL baseline model. It consists of assigning each word in the test data the majority class observed in the training data. The unseen words are given the tag 'O' (not a NE). The results obtained for the baseline as shown in Table 4 are a good indicator for the percentage of NE's already seen in the training phase.
- Parameter setting: we establish the impact of two experimental factors on NER performance, namely tokenization and the contextual window size as a preliminary pre-cursor to our feature engineering experiments. Clitic tokenization in a highly agglutinative language such as Arabic has been shown to be useful for many NLP applications [14]. Intuitively, clitic tokenization serves as a first layer of smoothing in such sparse high dimensional spaces. We empirically discover an optimal window size. We investigate window sizes of *-1/+1* to *-4/+4* tokens/words surrounding a target NE on the UPV-corpus. Table 3 shows the results obtained for the untokenized corpus (UNTOK) and the tokenized corpus (TOK), respectively. From Table 3 we note that clitic tokenization has a significant positive impact on NER. We see an increase of 3 absolute points in F1 score when the text is clitic tokenized. Moreover, a context size of *-1/+1* performs the best in this task. In fact, there seems to be a degrading effect correlated with window size, the bigger the window, the worse the performance.
- Feature engineering: we conduct different sets of experiments to explore the space of possible features. We use clitic tokenized text and we define the CXT to be *-1/+1* as established in the previous section. We explore individual features (always with CXT), combined features and then all the features together. We evaluate the performance in our experiments using 5-fold cross validation on each corpus independently. For the UPV-Corpus we have chosen the same ratio of test data size to training data size which has been used in the CoNLL competitions [22]. As far as the ACE data, we have replicated the same splits which were adopted in the ACE evaluations. Table 2 shows the average size of the training and test data for each corpus.

Table 3. Parameter setting experiments comparison between different window sizes and the impact of tokenization on the NER task.

	-1/+1	**-2/+2**	**-3/+3**	**-4/+4**
CXT+UNTOK	71.66	67.45	61.73	57.49
CXT+TOK	74.86	72.24	67.71	64

Table 4 illustrates the overall results yielded. We achieve state-of-art for almost all the corpora. We obtain an *F1* score up to 82.71 for ACE 2003, broadcast news genre. We note significantly improving over the baseline for all corpora with our best performing system all the features combined. Overall the worst results are yielded for the WL genre of data this may be explained by the overall randomness of the WL data relative to the other genres. The single best feature is the POS feature achieving an *F1* score of 78.97 on the UPV-Corpus and the highest performance for all the other corpora except for WL where the best single feature is MORPH feature. Combining all the features together yields the highest performance across the board for all corpora except the ACE 2005 BN corpus where the MORPH feature seems to reduce the combination performance (All except MORPH yields an F1 score of 82.13 compared to 81.47 yielded by the "All features" combination.

6. Discussion and Error Analysis

- CAP feature: using the CAP feature in our experiments shows whether the absence of capital letters in Arabic is a characteristic which hardens the NER task for the Arabic language or not. Table 4 shows that when the CAP feature is included the results improve significantly ranging from an increase in *F1* score of 0.5 to 4.5 absolute.
- MADA features: all the features we use in our experiments are language-independent except for the MORPH features. We recognize that a tool such as MADA (differently from the POS tagger and BPC chunker) is very language specific and expensive to produce for other languages. It is shown to have a very significant impact on more genres than others as illustrated by the significant improvement in NER on the WL ACE 2005 data.

Table 4. Feature engineering obtained results for different feature sets and their combinations.

	UPV-Corpus	ACE 2003		ACE 2004			ACE 2005		
	NW	BN	NW	BN	NW	ATB	BN	NW	WL
Baseline	31.5	74.78	69.08	62.02	52.23	64.23	71.06	58.63	27.66
CXT	74.68	72.76	68.27	70.74	63.16	63.26	74.53	65.5	34.53
CXT+LEX	77.18	73.22	72.46	71.02	63.6	63.43	76.32	67.18	30.8
CXT+GAZ	75.71	74.98	69.05	70.69	64.55	65.4	75.03	66.58	37.62
CXT+MORPH$_{raw}$	78.09	77.16	74.93	71.84	68.8	71.03	76.1	70.04	43.27
CXT+MORPH$_{mod}$	78.58	77.27	73.97	71.69	67.79	68.18	75.64	70.11	42.53
CXT+POS	78.97	78.48	74.44	73.75	69.86	72.58	77.28	70.67	35.04
CXT+BPC	76.18	72.9	70.7	71.3	64.03	64.99	74.77	65.17	31.63
CXT+NAT	75.71	74.19	70.68	71.8	63.86	65.02	75.04	66.62	36.66
CXT+CAP	79.13	76.78	72.31	72.03	65.1	64.01	74.9	67.54	36.47
CXT+POS+BPC	78.46	78.15	74.13	73.03	69.24	72.14	76.5	70.13	34.68
All except MORPH	79.86	81.87	78.31	75.06	71.93	73.78	82.13	75.91	42.67
All	80.4	82.71	79.21	76.43	73.4	75.34	81.47	76.19	53.81

Table 5. Results per class.

Class	Feature-set	UPV-corpus	ACE 2003		ACE 2004			ACE 2005		
		NW	BN	NW	BN	NW	ATB	BN	NW	WL
PER	All except MORPH	75.9	81.0	76.1	79.4	75.4	78.4	80.7	71.9	40.5
	All	78.5	81.4	77.0	80.0	77.4	78.0	80.7	73.3	56.1
LOC	All except MORPH	89.4	86.1	84.0	80.7	78.9	77.8	87.9	83.9	52.3
	All	89.6	87.0	84.4	82.6	80.1	80.5	87.2	84.2	58.9
ORG	All except MORPH	64.8	50.5	62.3	49.4	56.9	61.3	62.0	65.5	14.0
	All	64.3	51.3	65.6	50.7	58.3	63.4	60.9	65.0	23.8

Table 6. ACE 2004 BN errors confusion matrix.

	FAC	LOC	ORG	PER	VEH	WEA	O
FAC	34%	3.9%	9.1%	1.3%	0%	0%	51.7%
LOC	0%	81.82%	0.9%	0.9%	0%	0%	16.38%
ORG	0.3%	4.9%	49.8%	0.7%	0.2%	0%	44.1%
PER	0%	1.7%	0.2%	81.4%	0%	0%	16.7%
VEH	0%	2.2%	11.3%	13.6%	45.5%	6.8%	20.6%
WEA	0%	0%	0%	0%	0%	21.5%	78.5%

Table 4 shows the results of combining all the features except for MORPH and we note that there is a significant improvement over the singleton features or even combining two features such as POS and BPC. The addition of the MORPH features adds another boost for sure, but not as significant as the boost introduced by the overall joint combination in the modeling itself except in the WL genre.

In order to have a better idea on how the MORPH features help to capture the NEs we have added another table to show the results per class as shown in Table 5.

This table shows the results for each class separately for all the data-sets when the features sets 'All except MORPH' and 'All' were used. We report results only for the classes which are common to all the corpora, i.e., PERson (PER), LOCation (LOC) and ORGanization (ORG). As Table 5 shows, all the classes have benefited from the MORPH features. The PER class shows an improvement of 15 points on the WL genre.

According to our error analysis, those classes have benefited differently from the MORPH features.

Following, we give some examples for each of these classes:

- ORG: most of the ORG have a long span. Consequently, the NER model needs more information in order to be able to decide on the starting and ending words of an ORG NE. When we include the MORPH features, the model is able to obtain such information from the 'number' and 'gender'; i.e., if the elements of a word sequence have the same number and gender they have a high probability of belonging to the same class due to agreement properties. In the following examples, we underline the NEs which are captured after including the MORPH features: في هذه المنطقة على التصدي لحزب العمل الكردستاني transliterated as fy h*h AlmnTqp ElY AltSdy lHzb AlEml AlkrdstAny, translated as in this zone to confront the kurdistan labor party.

- LOC and PER: the following examples show NEs that were only identified and classified correctly after including the MORPH features. في مدينة مكسيكالي القريبة من الحدود ... Transliterated as: fy mdtnp mksykAly Alqrybp mn AlHdwd …Translated as:in the city of Mexicali which is close to the borders … كان دوينسبرج يتحدث قبل أن ... Transliterated as: wkAn dwynsbrj ytHdv qbl >n Translated as: and Dewinsberg was speaking before … we observe that most of the foreign NEs (non-Arabic words) are not tagged with a morphological analysis, MADA assigns them a "NO-ANALYSIS" tag. We use this as a strong signal that the word is a NE. Although the 'NO-ANALYSIS' tag is useful for most of the NE classes, it occurs more frequently with the LOC NEs.
- Class confusion: we examine a confusion matrix for each of the features-sets. An analysis of the confusion matrices shows that all the misclassified NEs are assigned the 'Outside' class. This result is another proof that the NER task is more complicated and challenging for the Arabic text because of its templatic morphology and the absence of capital letters as shown in section 3) which make the detection of the NEs existing in the text much harder. Table 6, shows the confusion matrix of the errors observed on the ACE 2004 BN data, using "All features". The rows show the reference tags and the columns the system output tags, i.e., a correct reading of the table would be, 34% of the FAC NEs are classified correctly by the system, 3.9% are misclassified as LOC, … 0.2% of the PER NEs are misclassified as ORG, *etc*.

7. Conclusions and Future Directions

We described a novel NER system using SVMs and a combination of both language independent and language dependent features for Arabic NER. We measure the impact of the different features independently and in a joint combination across different standard data sets and different genres. Our experiments yield state of the art performance significantly outperforming the baseline. Our best results achieve an *F1* score of 82.17 using all the features on the ACE 2003 BN data. Our results strongly suggest that employing language specific features in conjunction with language independent features is very beneficial for NER system performance. This result shows the relevance of morphological features for languages that exhibit complex and rich structures.

Furthermore, the features which we have used are all language-independent except the ones extracted with a morphological analysis and disambiguation (MADA) tool, which is able to extract up to fourteen morphological features, for each word. A significant improvement has been achieved when MADA was used for text in which the NE's appear in random contexts (i.e., weblogs). However, the use of MADA only helped to obtain an improvement of 0.8 points on average for the rest of the corpora. Some of these features, such as 'number' and 'gender' have been very helpful to capture long span NEs because they indicate to the model that a sequence of word share the same values of these features. On the other hand, they have also been very useful to capture the NEs which do not appear in the training data and are foreign to the Arabic language. Most of these words cannot be assigned any morphological analysis by MADA, and such information has been useful to detect the NEs, whereas the rest of the features and the contextual information have been used to identify the class of the NE.

For future work, due to the very positive impact noted for the use of gazetteer though they were of relatively small size, we plan to automatically extract bigger gazetteers using data mining techniques in order to enhance the performance of our system.

Acknowledgement

The authors would like to thank the reviewers for their detailed constructive comments. We would like to thank PCI-AECI A/010317/07, MCyT TIN2006-15265-C06-04 and TIN2009-13391-C04-03 research projects for partially funding this work. Mona Diab would like to acknowledge DARPA GALE Grant Contract No. HR001106-C-0023 for partially funding this work.

References

[1] Abuleil S., *Extracting Names from Arabic Text for Question Answering Systems*, Springer, 2002.

[2] Babych B. and Hartley A., "Improving Machine Translation Quality with Automatic Named Entity Recognition," *in Proceedings of EACL-EAMT*, Thailand, pp. 259-266, 2003.

[3] Benajiba Y., Rosso P., and José B., "ANERsys: An Arabic Named Entity Recognition System Based on Maximum Entropy," *in Proceedings of Conference on Computational Linguistics*, India, pp. 289-310, 2007.

[4] Benajiba Y. and Rosso P., "ANERsys 2.0: Conquering the NER Task for the Arabic Language by Combining the Maximum Entropy with POS-tag Information," *in Proceedings of Workshop on Natural Language Independent Engineering IICAI*, India, pp. 36-40, 2007.

[5] Bender O., Och J., and Ney H., "Maximum Entropy Models for Named Entity Recognition," *in Proceedings of the Conference on Natural Language Learning (CoNLL)*, Canada, pp. 142-147, 2003.

[6] Carbonell J., Harman D., Hovy E., Maiorano S., Prange J., and Sprack K., "Vision Statement to

Guide Research in Question and Answering and Text Summarization," *Report Technique NIST*, 2000.

[7] Chieu L. and Ng T., "Named Entity Recognition with a Maximum Entropy Approach," *in Proceedings of the Conference on Natural Language Learning (CoNLL)*, New York, pp. 160-163, 2003.

[8] Curran R. and Clark J., "Language Independent NER Uses a Maximum Entropy Tagger," *in Proceedings of the Conference on Natural Language Learning (CoNLL)*, UK, pp. 1-4, 2003.

[9] Diab M., Alkhalifa M., Elkateb S., Fellbaum C., Mansouri A., and Palmer M., "Semeval Task 18: Arabic Semantic Labeling," *in Proceedings of International Workshop on Semantic Evaluations (SemEval)*, Spain, pp. 81-86, 2007.

[10] Diab M., Moschitti A., and Pighin D., "CUNIT: A Semantic Role Labeling System for Modern Standard Arabic," *in Proceedings of RANLP*, USA, pp. 193-224, 2007.

[11] Diab M., Hacioglu K., and Jurafsky D., *Arabic Computational Morphology: Knowledge-based and Empirical Methods*, Springer, 2007.

[12] Farber B., Freitag D., and Habash N., "Improving NER in Arabic Using a Morphological Tagger," *in Proceedings of LREC,* Boston, pp. 96-99, 2008.

[13] Habash N. and Rambow O., "Arabic Tokenization, Part of Speech Tagging and Morphological Disambiguation in One Fell Swoop," *Workshop of Computational Approaches to Semitic Languages (ACL-2005*), NY, pp. 1-8, 2005.

[14] Habash N. and Sadat F., "Arabic Preprocessing Schemes for Statistical Machine Translation," *in Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL)*, Taiwan, pp. 209-219, 2006.

[15] Kudo T. and Matsumato Y., "Chunking with Support Vector Machine," *in Proceedings of the 4th Conference on Very Large Corpora*, Taiwan, pp. 18-25, 2000.

[16] Li W. and McCallum A., "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*, pp. 79-84, 2003.

[17] Maamouri M., Bies A., Buckwalter T., and Mekki W., "The Penn Arabic Treebank: Building a Large Scale Annotated Arabic Corpus," *in Proceedings of NEMLAR Conference on Arabic Language Resourcesand Tools*, Egypt, pp. 102-109, 2004.

[18] 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*, Canada, pp. 8-15, 1998.

[19] Malouf R., "Markov Models for Language Independent Named Entity Recognition," *in Proceedings of the Conference on Natural Language Learning (CoNLL),* Canada, pp. 30-34, 2003.

[20] Mayfield J., McNamee P., and Piatko C., "Named Entity Recognition Using Hundreds of Thousands of Features," *in Proceedings of the Conference on Natural Language Learning (CoNLL)*, Canada, pp. 68-72, 2003.

[21] Tran T., Thao X., Ngo H., Dinh D., and Collier N., "Named Entity Recognition in Vietnamese Documents," *Computer Journal of Progress in Informatics*, vol. 5, no. 3, pp. 14-17, 2007.

[22] Tjong F. and De Meulder F., "Introduction to the Shared Task: Language Independent Named Entity Recognition," *in Proceedings of the Conference on Natural Language Learning CoNLL-2003*, Canada, pp. 106-200, 2003.

[23] Toda H. and Kataoka R., "A Search Result Clustering Method Using Informatively Named Entities," *in Proceedings of the 7th ACM International Workshop on Web Information and Data Management*, USA, pp. 81-86, 2005.

[24] Vapnik V., *The Nature of Statistical Learning Theory*, Springer, 1995.

[25] Wu W., Jan Y., Tzong R., and Hsu L., "On Using Ensemble Methods for Chinese Named Entity Recognition," *in Proceedings of the 5th SIGHAN Workshop on Chinese Language Processing*, China, pp. 42-145, 2006.

[26] Zitouni I., Sorensen J., Luo X., and Florian R., "The Impact of Morphological Stemming on Arabic Mention Detection and Coreference Resolution," *in Proceedings of the Workshop of Computational Approaches to Semitic Languages, 43rd Annual Meeting of ACL*, France, pp. 79-86, 2005.

Yassine Benajiba has been granted a scholarship from the Spanish Agency of International Cooperation (AECI) in order to make PhD studies, Department of Informatics and Computation at EliRF, Polytechnic University of Valencia, Spain.

Mona Diab received her PhD in 2003 in the linguistics department from University of Maryland Institute for Advanced Computer Studies (UMIACS), University of Maryland College Park. Her PhD work focused on lexical semantic issues and was titled word sense disambiguation within a multilingual framework.

Paolo Rosso received his PhD in computer science in 1999 from the Trinity College Dublin, University of Ireland. He is currently an associate professor and the head of the Natural Language Engineering Laboratory at ELiRF, Polytechnic University of Valencia, Spain.