IEEE
IEEE Signal Processing Society
IEEE Transactions on
Audio, Speech and
Language Processing

Arabic Named Entity Recognition: A Feature-driven Study

Journal:	*Transactions on Audio, Speech and Language Processing*
Manuscript ID:	T-ASL-02013-2008.R1
Manuscript Type:	Processing Morphologically Rich Languages
Date Submitted by the Author:	n/a
Complete List of Authors:	Benajiba, Yassine; UPV, DSIC Diab, Mona; Columbia University, CCLS Rosso, Paolo; UPV, DSIC
EDICS:	SLP-UNDE Spoken Language Understanding < SPOKEN LANGUAGE PROCESSING, SLP-SMIR Speech Data Mining and Document Retrieval < SPOKEN LANGUAGE PROCESSING, SLP-LANG Language Modeling (for Speech and SLP) < SPOKEN LANGUAGE PROCESSING

scholarONE™
Manuscript Central

Only

Arabic Named Entity Recognition: A Feature-driven Study

Yassine Benajiba, Mona Diab and Paolo Rosso

Abstract—**The Named Entity Recognition (NER) task aims at identifying and classifying Named Entities within an open-domain text. This task has been garnering significant attention recently as it has been shown to help improve the performance of many natural language processing (NLP) applications. In this paper, we investigate the impact of using different sets of features in three discriminative machine learning frameworks, namely, Support Vector Machines, Maximum Entropy and Conditional Random Fields for the task of NER. Our language of interest is Arabic. We explore lexical, contextual and morphological features and nine data-sets of different genres and annotations. We measure the impact of the different features in isolation and incrementally combine them in order to evaluate the robustness to noise of each approach. We achieve the highest performance using a combination of fifteen features in Conditional Random Fields using Broadcast News data ($F_{\beta=1}$=83.34).**

Index Terms—**Arabic, Natural Language Processing, Named Entity Recognition, Machine Learning Comparison.**

For Review Only

I. 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 (NEs) within an open-domain text. For instance, for the following text:

‘John left Washington D.C. at 4 a.m. and reached New York City at 7h30 a.m.’

A NER system should be able to identify ‘John’, ‘Washington D.C.’ and ‘New York City’ as NEs, and classify the first one as a person and the second and third ones as locations. 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.

We list here some NLP systems that benefit from NER:

- Information Retrieval (IR): The IR system’s main goal is to retrieve, from a large document-set, the relevant documents to the user query. In [Thompson and Dozier1997], the authors carried out a statistical study of the percentage of user queries, over a period of several days, to different news databases containing NEs. The authors report that 67.83%, 83.4% and 38.8% of the queries contained a NE in the Wall St. Journal, Los Angeles Times and Washington Post, respectively. In the same paper, the authors report the results obtained when a NER system was integrated in the IR system in order to consider each NE as a single term when the IR system computes the *tf-idf* [Salton and Buckley1988] (term frequency-inverse document frequency). The NE-based approach outperformed the baseline precision (a probabilistic retrieval engine [Turtle and Croft1991]) on all recall levels.
- Question Answering (QA): QA systems aim at answering user specific questions with accurate answers. In [Ferrández et al.2007] the authors argue that a study of the percentage of the questions containing one or more named entities in the CLEF[1] 2004 and 2005 competitions, showed that the majority, precisely 87.7%, of questions contained a NE. Moreover, in [Greenwood and Gaizauskas2007] the authors state that the performance of a QA system can be considerably improved if a NER system is used for a question whose answer is a NE. In their work, the authors show that using an accurate NER system has helped improve the ratio of correctly answered questions of the type *‘When did X die?’* from 0% to 53%.
- Machine Translation (MT): The translation of NEs requires different approaches than the translation of common words. For instance, ‘Beijing’ should be translated as ‘Pekin’ in French, whereas ‘Paris’ should remain unchanged. Other NEs need a character-based transliteration such as ‘Hizbullah’ or ‘Ghandi’. In [Babych and Hartley2003], the authors show that a considerable error rate is experienced when the automatic translator does not manage to transliterate properly the NEs. In order to study the possibility of improving the performance of a MT system by embedding a NER system, they have tagged all the text which has to be translated by a NER system as a pre-processing step. Thereafter, the words tagged by the NER system were translated using the methods which are specific for NEs translation. The results showed that this technique out-

Y. Benajiba is a Ph.D. student in the Department of Informatic Systems and Computation at the Polytechnic University of Valencia and a member of the Natural Language Engineering (NLE) Lab. Camino de Vera, s/n, 46022, Valencia, Spain. email: benajibayassine@gmail.com, URL: http://www.dsic.upv.es/∼ybenajiba

M. Diab is an Associate Research Scientist in the Center for Computational Learning Systems at Columbia University. 850 Interchurch Center / MC 7717, 475 Riverside Drive, New York, NY 10115, USA. email: mdiab@ccls.columbia.edu, URL: http://www.cs.columbia.edu/∼mdiab/

P. Rosso is a permanent lecturer in the the Department of Informatic Systems and Computation at the Polytechnic University of Valencia and head of the Natural Language Engineering (NLE) Lab. Camino de Vera, s/n, 46022, Valencia, Spain. email: prosso@dsic.upv.es, URL: http://www.dsic.upv.es/∼prosso

[1] http://www.clef-campaign.org

performs the methods which do not consider tagging the NEs before the translation.

- Text clustering: Search result clustering (a sub-task of Text Clustering) is the NLP task focused on clustering in groups the results returned by a search engine. For instance, if we have the documents returned by a search engine for the query *'Michael Jordan'*, in order to make these results easier to explore for the user, a result clustering system would cluster the documents concerning Michael Jordan the basketball player[2] in one cluster, and the ones relevant to the Berkeley professor[3] in another cluster. In [Toda and Kataoka2005], the authors report that they have outperformed the existing search results clustering by including a NER system in their global system as it attributes a special weight to the NEs in their clustering approach.

In this paper, we consider the problem of NER for Arabic[4]. The NER task in morphologically rich languages such as Arabic is relatively different from performing the task in English due to inherent characteristic linguistic differences, such as the agglutinative nature of the language allows for complex and hence more sparse data representation. Orthographic characteristics such as lack of capitalization to mark a named entity also render the NER task more challenging in Arabic. We compare between three discriminative approaches to the NER problem: Support Vector Machines (SVMs)[Vapnik1995], Maximum Entropy (ME)[Berger et al.1996] and Conditional Random Fields (CRFs)[Lafferty et al.2002]. We comprehensively investigate many sets of features: contextual, lexical, morphological and shallow syntactic features. We explore the features in isolation first. Thereafter we rank the features according to their impact and we evaluate the approaches using the N features with the highest impact. We have conducted experiments for *N=1* to *N=total number of features* in order to show the robustness to noise of each of the mentioned approaches. We experiment with two sets of data, the standard ACE data and a manually created data set, NLE-corpus, in order to confirm the reliability of our results. Our best system that combines the fifteen best features using CRFs yields an overall F1 score of 83.34.

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

II. Related Work

There are several significant research efforts in NER. In the Conference on Natural Language Learning (CoNLL) 2003 NER evaluation tasks[5] the participants were asked to identify the NEs within the test-sets (English and German) and classify them. The system which has yielded the best performance is described in [Florian et al.2003]. The authors report that they have used a linear interpolation of three different classifiers: (i) Hidden Markov Models; (ii) Maximum Entropy (ME); and (iii) Robust Risk Minimization (RRM). Their final results were 88.76 for English and 72.41 for German, best results in both languages. [Chieu and Ng2003] was ranked the second best participation. The system is fully ME-based and uses different types of features, namely, contextual and lexical features as well as capitalization. The authors performed two runs where the second one uses additional external resources. For English, the results using the lexicon (88.12) were almost two points higher than those obtained without using an external resource (86.83). However, the results were higher when no external resource was used for German (77.05 vs. 76.83). Finally, the system of [Klein et al.2004] was ranked third. It employed a character-based HMM approach. This method relies heavily on internal evidence for the NEs. Their final results for English were 86.07 (third best results) and for German 71.9 (second best results).

[Tran et al.2007] show that using a Support Vector Machine (SVM) approach outperforms ($F_{\beta=1}$=87.75) using CRFs (86.48) on the NER task in Vietnamese. However, this comparison is based on the average F-measure obtained by using the same feature-set with both SVMs and CRFs. Thus it is not possible to make any further conclusions on the behavior of these Machine Learning (ML) approaches with different features. In [Zhang and Johnson2003], the authors report a manual feature selection study in which they prove that simple token-based features can be more helpful than sohpisticated linguistic features. In this study, we can find the performance obtained with different feature-sets, however the authors did not compare the performance of the different ML approaches.

With current surge in resources making their way in the NLP community for Arabic, we are starting to see systems being developed for the processing of the Arabic language.

In work by [Zitouni et al.2005], the authors use a 2-step approach (NE boundary detection and then NE classification) in their investigation of Arabic mention detection problem. A mention refers to a named entity (e.g. Ohio), a nominal (e.g. Prime Minister), or a pronominal (e.g. he) entity. They pre-process the data applying morphological stemming. They adopt an ME Markov model approach exploring lexical, syntactic and gazetteer features. The authors evaluate their system's performance against the Automatic Content Extraction (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, i.e. we are not able to tell the performance on NER task specifically. On the task of Arabic NER alone, there has been recent significant work. In some of our recent work, [Benajiba et al.2007], we show that using a basic ME approach to Arabic NER yields an F1-measure of 55.23. We

[2] http://en.wikipedia.org/wiki/Michael_Jordan

[3] http://www.cs.berkeley.edu/∼jordan/

[4] We use Arabic in this paper to refer to Modern Standard Arabic.

[5] http://www.cnts.ua.ac.be/conll2003

For Review Only

For Review Only

evaluate the system using a corpus that was developed in-house using CoNLL guidelines, NLE-corpus. We followed up with further work in [Benajiba and Rosso2007] which yields results reaching $F_{\beta=1}$=65.91 by adopting a two stage classification using an ME based approach to the problem in a style similar to [Zitouni et al.2005]. Finally, in our most recent work on the problem, we explore using CRFs in [Benajiba and Rosso2008] with different features, namely, contextual, morphological, and lexical features, together with gazetteers as external resources. We report the performance obtained with each feature in isolation and our best results were acheived when all the features were combined (79.21). Finally, [Farber et al.2006] use a structured perceptron as well as sophisticated morphological features for the task of Arabic NER. The authors report achieving an F-measure of 75.7 on the newswire subset of the ACE 2005 corpora.

III. Arabic in the context of Named Entity Recognition Task

Let us consider the example presented in Figure 1. In Buckwalter transliteration it can be written as *whw llsnp bmvAbp Alqmr ynyr lylA lyzyn AlsmA'*[6]. The underlined word ('llsnp') can be read as 'to the year' or 'to the Sunnah', i.e. in the latter case it is an NE. If we consider that the first case to be correct then the whole sentence can be translated as 'and it is to the year as the moon which lightens in the night to beautify the sky', and 'it' could refer to a special month or day in the year. In the second case, the correct translation would be 'and he is to the Sunnah as the moon which lightens in the night to beautify the sky', and 'he' could refer to someone of great importance for the 'Sunnis'. Hence, if we do not have any other information it would not be possible to disambiguate whether 'llsnp' is a NE or not for the two following reasons:

وهو للسّنة بمثابة القمر ينير ليلا ليزين السماء

Fig. 1. An illustrating example of the difficulty of Arabic NER

- *Absence of short vowels:* In newpaper articles, magazines, books and all resources written in MSA, the texts are mostly unvocalized. The vocalized form of the word 'llsnp' is spelled differently, 'llsunnap' (meaning 'for the Sunnah') vs. 'llsanap' (meaning 'for the year') disambiguating between the two readings. Human readers use context and knowledge in order to disambiguate the unvocalized forms. However, in the case of an NER system, the considered context is typically limited and the knowledge (lexicons, gazetteers, etc.) sources are inherently static and limited in scope.
- *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. Hence, in our example there is no way we can capitalize (or mark) the NE given the Arabic orthography rules.
- *Sparseness:*From a statistical NLP viewpoint, morphologically rich languages have another more important obstacle generally known as 'data sparseness'. These languages use an agglutinative strategy to form surface tokens. In the case of Arabic, being a Semitic language,[7] it exhibits a templatic morphology where words are made up of roots and affixes. Clitics agglutinate to words. For instance, the surface word in Figure 2 *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'.

وبحسناتهم

Fig. 2. An illustrating example of the templatic morphology of the Arabic language

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 (from a statistical viewpoint) is much more sparse compared to English. In order to tackle this problem, we need to perform a level of *segmentation of the clitics for each word* (*clitic-segmentation*) as a pre-processing step. It is particularly useful for NER as: (i) the NEs always appear in the same form hence lowering the number of unseen NEs; (ii) it reduces the number of different surface form contexts in which the NEs appear.

IV. Approach using A Large Range of Features

A. The SVMs, ME and CRFs Approach

In this paper, we explore three different yet comparable approaches that were used for NER in other languages: SVMs, ME, and CRFs. The approaches have well known desirable characteristics for NLP applications.

SVMs are proven to be robust to noise and to have a 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 [Diab et al.2004], [Kudo and Matsumato2000], and for the NER task in particular [Mayfield et al.2003], [Tran et al.2007]. In order to use SVMs in the NER task, we use *Yamcha*[8] toolkit.

ME aims at providing a model with the less biases possible [Berger et al.1996]. One of the first implementations of the ME approach in NLP tasks is [Ratnaparkhi1996]. Moreover, as mentioned in section II, this approach proved to be successful for the NER task. We implemented our own ME approach to carry out the experiments, for weight estimation we use *Yasmet*.[9]

CRFs are oriented to segmenting and labeling sequence data [Lafferty et al.2002]. As undirected graphical models, CRFs

[6] For purposes of presentation, we adopt a Buckwalter transliteration scheme to show romanized Arabic [Buckwalter2002].

[7] Other Semitic languages include Hebrew and Amharic

[8] http://chasen.org/∼taku/software/yamcha/

[9] http://www.fjoch.com/YASMET.html

are alternative ways to represent probability distributions. During the training phase the conditional probabilities of the classes are maximized. CRFs are proven to be very efficient for the NER task (see section II). We use *CRF++*[10] for our experiments.

B. Arabic NER Task Tag sets

While different NLP applications may require different tag sets, we address here the NER task as an end system in itself. There exist three standard NER tag sets in the literature:

(i) Message Understanding Conference (MUC-6)[11]: the NER task consisted of three subtasks:

- ENAMEX: for proper nouns;
- NUMEX : for numerical expressions; and
- TIMEX : for temporal expressions.

The ENAMEX subtask was defined as the identification of the NEs and their classification:

- Person, e.g. Albert Einstein;
- Location, e.g. Paris;
- Organization, e.g. Google Co.

(ii) Conference of Natural Language Learning (CoNLL): In the language-independent NER shared task held in the CoNLL 2002[12] and CoNLL 2003[13] the tag-set comprised four classes: Person, Location, Organization (same as previous ones) and *Miscellaneous (e.g. Empire State building)*;

(iii) ACE: The ACE 2003 data defines four different classes: Person, Geographical and Political Entities (GPE), Organization and Facility. Whereas in ACE 2004 and 2005 two classes were added to the previous 2003 tag set: *Vehicles (e.g. Rotterdam Ship)* and *Weapons (e.g. Kalashnikof)*.

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.

C. Features

The most challenging aspect of any machine learning approach to NLP problems is deciding on the optimal feature sets. In this work, we investigate a large space of features. The feature sets are 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_i):** This feature defines the lexical orthographic nature of the tokens in the text. We define it as a character n-gram of 6 characters. This feature is elaborated in the following example. Consider that a word is simply a sequence of characters $C_1C_2C_3...C_{n-1}Cn$ then the lexical features would be

- LEX_1=C_1
- LEX_2=C_1C_2
- LEX_3=$C_1C_2C_3$
- $LEX4$=C_n
- $LEX_5 = C_{n-1}C_n$
- $LEX_6 = C_{n-2}C_{n-1}C_n$

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[14] as well as other web sources. This enrichment consisted of: (i) taking the page labeled '*Countries of the world*' (dwl AlEAlm) as a starting point to crawl into Wikipedia and retrieve location names; (ii) we automatically filter the data removing stop words; (iii) finally, we manually filter the resulting set ensuring its good quality as a source of location names.

Morphological features (M_x): This feature set is based on exploiting the rich characteristic morphological features of the Arabic language. We relied on a system for Morphological Analysis and Disambiguation for Arabic (MADA) to extract relevant morphological information [Habash and Rambow2005]. 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 raw 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-segmentation.

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. Although MADA produces fourteen features we only use eleven of them, the description of each one of these features goes beyond the scope of this paper, yet they are explained in detail in [Habash and Rambow2005]. However, the eleven features used in our experiments are listed as follows:

- M_{ART}=article: indicates whether a token has a definite article or not;
- M_{ASP}= verb aspect: In Arabic, a verb maybe imperfective, perfective or imperative.
- M_{CASE}=grammatical case: genitive, accusative, nominative;
- M_{CLIT}=clitic: indicates whether a word has any clitics attached;
- M_{CONJ}=conjunction: MADA indicates whether a token has any conjunction attached or not;
- M_{DEF}=definiteness: MADA indicates whether a token is definite or not. All the NEs by definition are definite;
- M_{MOOD}=mood: indicative, imperative, subjuntive and optative are the possible values of this feature;

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

[11] http://cs.nyu.edu/cs/faculty/grishman/muc6.html

[12] http://www.cnts.ua.ac.be/conll2002/

[13] http://www.cnts.ua.ac.be/conll2003/

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

For Review Only

- M_{NUM}=number: For almost all the tokens categories (verbs, nouns, adjectives, etc.) MADA provides the grammatical *number*. In Arabic, the possible values are singular (SG), dual (DU) and plural (PL);
- M_{PART}=particle: MADA indicates whether a token has any particles attached or not;
- M_{PER}=person: In Arabic, verbs, nouns, and pronouns typically indicate person information. The possible values are *first, second* or *third* person;
- M_{VCE}=voice: In Arabic, a verb can have one of two possible voice values: active or passive.

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[15] described in [Diab et al.2007]. Like the MADA system, AMIRA-1.0 is an SVM based set of tools. The POS tagger performs at 96.2% and the BPC system performs at 96.33%. 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 POS tag set of 25 tags created for parsing and included in the Arabic Treebank [Maamouri et al.2004] distribution.

Nationality (NAT): We mark nationalities in the input text. Such information is useful for detecting NEs since they are used as precursors to recognize NE. Especially, when location and person NEs are introduced in a text they are usually preceeded by a nationality. For instance, الرئيس **الامريكي** جورج بوش, which can be trasliterated as “Alr¿ys **AlAmryky** *jwrj bw$*”, and translated to English as “The **American** President *George Bush*”. Another example is العاصمة **الاسبانية** مدريد, which can be transliterated as “AlEAsmp **AlASbanyp** *mdryd*” and translated to English as “the **Spanish** capital *Madrid*”.

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. The insight is that if the translation begins with a capital letter, then it is most probably a NE.

V. Experiments and Results

A. Data

We use the ACE 2003, 2004 and 2005 corpora and an enhanced version of the corpus used in [Benajiba et al.2007], NLE-corpus. Table I describe for the different corpora: the training data size ($Size_{train}$), the test size ($Size_{test}$).

$Corpus$	$genre$	$Size_{train}$	$Size_{test}$
NLE-corpus	NW	144.48k	30.28k
ACE 2003	BN	16.34k	2.51k
	NW	29.44k	7k
ACE 2004	BN	50.44k	13.32k
	NW	51.74k	13.4k
	ATB	21.27k	5.25k
ACE 2005	BN	22.3k	5k
	NW	43.85k	12.3k
	WL	18k	3.2k

TABLE I
Characteristics of NLE-Corpus and ACE 2003, 2004 and 2005 data

NLE-corpus

This corpus (see Table I) 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 CoNLL tag set as described in Section IV-B. The CoNLL tag set comprises 4 classes: Person, Location, Organization and Miscellaneous. The annotators followed the IOB2 annotation guidelines [Ratnaparkhi1996]. A NE can comprise more than one token. The IOB2 annotation scheme tags the first token in a NE as $B-Class$, $I-Class$ for each subsequent token, and O for the tokens which are not part of any NE. For instance:

‘John left Washington D.C. at 4 a.m. and reached New York City at 7h30 a.m.’

the IOB2 annotation scheme tags it as:

‘John/**B-PER** left Washington/**B-LOC** D.C./**I-LOC** at 4 a.m. and reached New/**B-LOC** York/**I-LOC** City/**I-LOC** at 7h30 a.m.’

ACE data

The ACE data (see Table I) is annotated for many tasks: Entity Detection and Tracking (EDT), Relation Detection and Recognition (RDR), Event Detection and Recognition (EDR). The ACE 2003 comprises two data genres: *Broadcast News* (BN) and *Newswire* (NW). The ACE 2004 additionally comprises the *Arabic Treebank* (ATB). The ACE 2005 does not have the ATB genre but it has *Weblogs* (WL). The ACE data annotates pronominal, nominal and named mentions of entities, since it caters to more than one type of task. For our purposes, we specifically care about the NER aspect of the data, hence, we discard the annotations of pronominal and nominal mentions and keep only the named mentions.

The only main difference which remains between the ACE annotation and the NLE-corpus ones is that in the former, nationalities (e.g. Spanish, French, etc.) are tagged as geopolitical entities (GPE) whereas in the ACE annotation scheme they are not considered NEs.

B. Experimental Set-up

1) Metrics: We use the CoNLL[16] evaluation standard metrics of precision, recall and F1-measure, which is the harmonic mean between precision and recall.

The CoNLL evaluation metrics are aggressive metrics in that they do not assign partial credit. A NE has to be identified as a whole (full span) and correctly classified in order to obtain credit.

[15] http://www1.cs.columbia.edu/∼mdiab/software/AMIRA-1.0.tar.gz

[16] http://www.cnts.ua.ac.be/conll2003/

For Review Only

For Review Only

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

a- 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 (see Table IV) indicate the percentage of NEs already seen in the training phase.

b- Parameter setting:

We need to establish the impact of two experimental factors on NER performance, namely clitic-segmentation and contextual window size as a preliminary pre-cursor to our feature engineering experiments. Clitic-segmentation in a highly agglutinative language such as Arabic has been shown to be useful for many NLP applications [Habash and Sadat2006]. Intuitively, clitic-segmentation serves as a first layer of smoothing in such sparse high dimensional spaces. We need to decide on an optimal window size, so we experiment with different sizes. We set the clitic-segmentation to the ATB standard clitic-segmentation scheme.[17] In these experiments, we investigate window sizes of $-1/+1$ up to $-4/+4$ tokens/words surrounding a target NE. We carry out the experiments on the NLE-corpus using SVMs without any additional features. Table II shows the CoNLL results obtained for the raw text corpus (UNSEG) and the segmented clitics corpus (SEG), respectively.

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

TABLE II
PARAMETER SETTING EXPERIMENTS: COMPARISON BETWEEN DIFFERENT WINDOW SIZES, AND THE IMPACT OF CLITIC-SEGMENTATION ON THE NER TASK

From Table II we note that clitic-segmentation has a significant positive impact on NER. We see an increase of 3 absolute points in F1 score when the text is clitic segmented. 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.

c- Feature engineering:

We conduct different sets of experiments to explore the space of possible features. We use clitic segmented text and we define the context (CXT) to be $-1/+1$ as established in the previous section. The rest of our experiments are organized as follows:

1) Explore individual features[18]: which consists of measuring the impact (number of F-measure points of improvement obtained) when each of the feature is used separately using each of the ML approaches and data-sets which we have described earlier;
2) Rank features according to their impact: from the obtained results in the previous step, we get a ranking of the features for each ML approach and data-set. In this step we aim at deducing a general ranking of the features. The algorithm which we have used for this purpose, assigns to each feature the most frequent rank. Table III illustrates the ranking obtained in our experiments.
3) Evaluate SVMs, ME and CRFs approaches combining each time the top N-top elements of the ranked features list. We have carried out experiments starting from N*=1* and up to from N*=22* to find the optimal number of features.

Rank	Feature	Rank	Feature
1	POS	12	NAT
2	CAP	13	LEX_1
3	M_{ASP}	14	LEX_4
4	M_{PART}	15	M_{CASE}
5	LEX_6	16	M_{NUM}
6	LEX_3	17	M_{DEF}
7	M_{CLIT}	18	LEX_2
8	BPC	19	LEX_5
9	GAZ	20	M_{CONJ}
10	M_{ART}	21	M_{MOOD}
11	M_{VCE}	22	M_{PER}

TABLE III
FEATURES RANKED ACCORDING TO THEIR IMPACT.

We evaluate the performance in our experiments using 5-fold cross validation on each corpus independently. For the NLE-corpus we have chosen the same ratio of test data size to training data size which has been used in the CoNLL competitions. For the ACE data, we have replicated the same splits which were adopted in the ACE evaluations. (Table I shows the average size of the training and test data for each corpus).

Figure 3 shows the results for ACE 2003 data, BN genre (best results). Figure 4 illustrates the results for the ACE 2004 data. Figure 5 shows the results obtained with ACE 2005, WL genre (worst results). Finally, table IV presents the baseline and the best results obtained for each corpus together with the number of features N and the corresponding ML approach. In the same table we also present the results which were obtained when all the features were combined.

VI. DISCUSSION AND ERROR ANALYSIS

We achieve state-of-art and significantly improve over the baseline for almost all the corpora. We have obtained an F1 score up to 83.34 for ACE 2003, Broadcast News genre. The worst results are yielded for the WL genre of data which may be explained by the overall randomness of the WL data relative to the other genres, in addition to the fact that WL data includes dialectal language which plays havoc with the basic processing tools such POS tagging and morphological disambiguation.Farber et al., [Farber et al.2006], report an F-measure of 75.7 on the NW genre of the ACE 2005 data using a set of morphological features. Our approach outperforms

[17] The ATB clitic-segmentation scheme consists of separating from the stem word: (i) prefixes are typically conjunctions and preposition; and (ii) pronominal suffixes.

[18] Due to space limitations, the results are not presented in this paper.

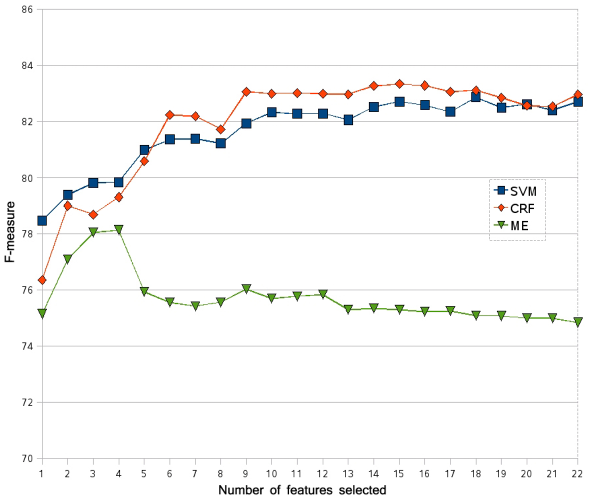

Fig. 3. Results per approach and number of features for the ACE 2003 (Broadcast News genre) data.

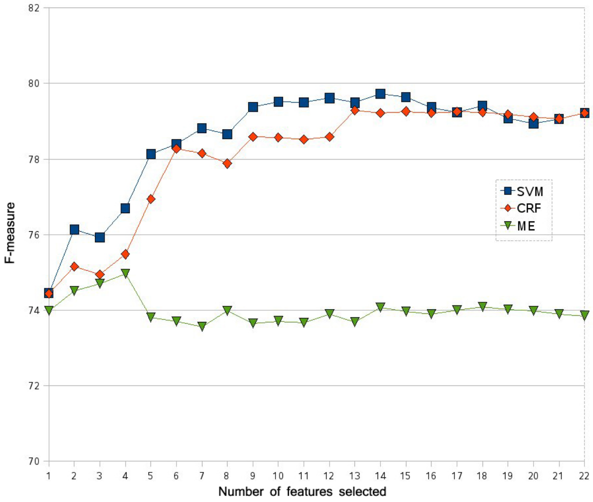

Fig. 4. Results per approach and number of features for the ACE 2003 (Newswire genre) data.

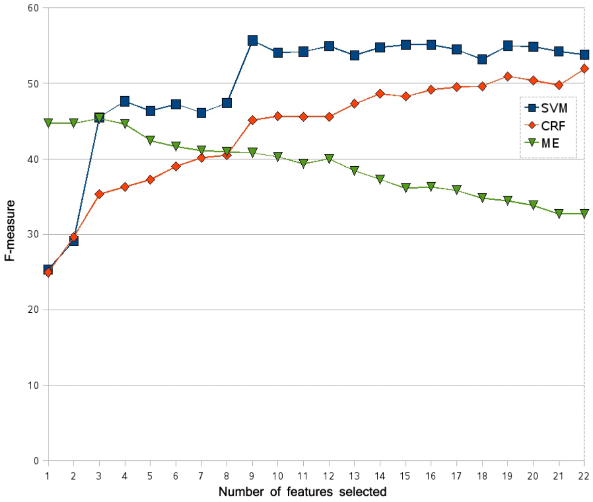

Fig. 5. Results per approach and number of features for the ACE 2005 (Weblogs genre) data.

Corpus	genre	Baseline	Best						All Features		
			SVMs		ME		CRFs		SVMs	ME	CRFs
			N	F-score	N	F-score	N	F-score			
NLE-corpus	NW	31.5	14	**81.04**	3	77.9	12	80.36	80.4	76.8	79.8
ACE 2003	BN	74.78	15	82.72	3	78.05	15	**83.34**	82.71	74.84	82.94
	NW	69.08	14	**79.72**	3	74.56	13	79.52	79.21	73.84	79.11
ACE 2004	BN	62.02	16	**77.61**	2	73.34	13	77.03	76.43	69.44	76.96
	NW	52.23	14	74.13	3	68.13	12	**74.53**	73.4	63.13	73.47
	ATB	64.23	15	75.43	2	69.95	13	**75.51**	75.34	64.66	75.48
ACE 2005	BN	71.06	15	**82.02**	3	77.67	14	81.87	81.47	75.71	81.1
	NW	58.63	15	76.97	3	70.31	13	**77.06**	76.19	67.41	75.67
	WL	27.66	12	**55.69**	2	44.96	14	53.91	53.81	32.66	51.81

TABLE IV
BEST OBTAINED RESULTS FOR EACH CORPUS.

theirs, yielding an F-measure of 77.06 (i.e. 1.36 points absolute difference) for the same data-set by using CRFs with the 13 best features (see Table III).

Features:

Let us consider the following sentence:

الرئيس الروسي فلاديمير ب # وطن في اقرب ...

in Buckwalter transliteration as:

Alrys Alrwsy flAdymyr b# wTn fy Aqrb ...

tranlated to English as:

The Russian president Vladimir Putin in the nearest ...

The word ’Putin’, which in Arabic is generally spelled بوتن (*bwtn*) is spelled differently in that specific text as بوطن(*bwTn*). Moreover, the clitic segmenter mistakenly split the first character from the word treating it as a prepositional prefix ب, meaning ’in’ or ’with’. Furthermore, the wrong spelling of *wTn* confuses the system further with the word meaning country which has the same spelling. Hence the word (*bwTn*) is misclassified as an O. Even if a NE can accept a prefix, it attaches to the first token in the NE.

The CAP feature is ranked second (see table III) among all others. This result confirms that the lack of capitalization in some languages such as Arabic considerably complicates the NER task. The use of lexical features (LEX_i) shows that only marking the first and last three characters of a word (LEX_3 and LEX_6) can be useful for a NER approach. The rest of the lexical features occur randomly with all the classes. The lexical features are mostly useful when the same NE appears slightly different in the different parts of the corpus. For instance, in the sentence:

تقدم ايريل شارون ...

transliterated as:

tqdm Ayryl $Arwn ...

translated to English as:

Ariel Sharon presented ...

The name ‘Ariel’ is transliterated in Arabic as اريل (*Aryl*) or ايريل (*Ayryl*). In the training corpus, this name appears only with the first transliteration. Hence, when seen in the latter spelling form in the test data, the classifier assigns it an O tag. However, when we employ the last trigram of the word as a feature, (LEX_6), the classifier correctly classifies the alternate spelling, ايريل (*Ayryl*) as part of a NE. It shares the same last three characters with the word اريل (*Aryl*) which has been frequently seen as a person in the training data. Another similar example is the NE الواشنطن بوست (*AlwA$ntn bwst*) ‘The Washington Post’ which appears with the definite article (*Al*) only once in the corpus. The classifier tags the word correctly only when the lexical feature LEX_6 is used.

On the other hand, the lexical feature LEX_3, which concerns the first three characters of each word, is mostly useful for NEs with different suffixes. The most remarkable example of such NEs are the nationalities which are tagged (depending on the context) as LOC, PER or O. Similar to English the difference between plural and singular forms of most of the nationalities is the suffix (e.g. ‘Palestinian’, فليسطيني (*flsTyny*) vs. ‘Palestinian**s**’, فليسطيني **ين** (*flsTyny***yn**). In addition, the Arabic has the dual form which is very rarely used but also requires only adding a suffix to the singular form. Hence ignoring the suffixes and focusing on first 3 letters in a token works in our favor in the case of nationalities. In our data, LEX_i features are very useful in capturing those cases.

Incremental Features Selection: The incremental feature selection yields slightly better results than using all features together. Moreover, it is important to notice that the time to extract, train and test with only 14 or 15 features is almost half the time necessary for 22 features. Through the examples of errors corrected when the best feature-set is used, we note that simply when we use a selected feature-set, we avoid providing the classifier noisy information. One case is the NE ‘Holy Shrine’, a facility FAC, in Arabic الحرم القدسي (*AlHrm Alqdsy*) is correctly classified with the best feature set. If all the features were to be used, for example the starting bigram, the definite article ’Al’, the classifier mis-tags the tokens as O.

Approaches: The results obtained with ME are considerably lower than the ones obtained by CRFs and SVMs

For Review Only

For Review Only

especially when the number of features exceed 6. This shows that the ME approach is much more sensitive to noise and that it is more suitable to use this approach when a restricted number of accurate features is used. On the other hand, CRFs and SVMs show very similar performance. Even though SVMs show a slightly better performance when only the first 7 top features are used. Thus, according to our results it is not possible to determine an absolute superiority of the SVMs or the CRFs for the Arabic NER task. Through the data we have also observed that even if SVMs and CRFs give different 'false alarms' they tend to miss the same NEs. Hence there is no real complementarity between the approaches.

Accordingly, the choice of one or the other has to be based on the number of available features and their quality.

VII. Conclusions and Future Directions

We describe the performance obtained using language-dependent and language independent features in SVMs, ME and CRFs for the NER task on different Arabic data-sets of different genres. We measure the impact of each feature individually, we rank them according to their impact and then perform incremental features' selection considering each time the N best features (we exhaustively explore all the possible values of N). Our experiments yield state of the art performance significantly outperforming the baseline. Our best results achieve an F1 score of 83.34 using the 15 best features in the CRFs approach on the ACE 2003 BN data. Our results show that the SVMs and CRFs have very similar behaviors and significantly outperform the ME approach. They also strongly suggest that the choice of using the CRFs or the SVMs approach should be based on the number of features available, i.e. if only few features are avaailable it is more suitable to use SVMs. Using the Arabic language in our experiments showed that a better performance is obtained in the NER task for languages which exhibit complex and rich structures if the data is pre-processed by a clitic-segmenter. They also show these languages can profit from their morphological richness if morphological features for each word are extracted and provided to the classifier. Those morphological features help the classifier by indicating that a word has a certain number of characteristics which makes it more or less probable to be a NE.

Other features such as the lexical features, which are based on the starting and ending character trigrams of the word, are totally language-independent, very easy to extract and they show that they can be very useful to capture NEs which might appear with a slight difference in the surface form in their occurrences in the data.

Acknowledgments

Acknowledgments here

References

[Babych and Hartley2003] Bogdan Babych and Anthony Hartley. 2003. *Improving Machine Translation Quality with Automatic Named Entity Recognition*. In *Proc. of EACL-EAMT*. Budapest.

[Benajiba and Rosso2008] Yassine Benajiba and Paolo Rosso. 2008. *Arabic Named Entity Recognition using Conditional Random Fields.* In *In: Proc. Workshop on HLT NLP within the Arabic world. Arabic Language and local languages processing: Status Updates and Prospects, 6th Int. Conf. on Language Resources and Evaluation, LREC-2008.*

[Benajiba et al.2007] Yassine Benajiba, Paolo Rosso and José Miguel Benedí. 2007. *ANERsys: An Arabic Named Entity Recognition system based on Maximum Entropy.* In *Proceedings of 8th Int. Conf. on Comput. Linguistics and Intelligent Text Processing, CICLing-2007*, Springer-Verlag, LNCS(4394), pp. 143-153.

[Benajiba and Rosso2007] Yassine Benajiba and Paolo Rosso. 2007. *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, 3rd Indian Int. Conf. on Artificial Intelligence, IICAI-2007.*

[Bender et al.2003] Oliver Bender, Franz Josef Och, and Hermann Ney. 2003. *Maximum Entropy Models For Named Entity Recognition.* In *Proceedings of CoNLL-2003*. Edmonton, Canada.

[Berger et al.1996] Adam L. Berger, Vincent J. Della Pietra and Stephen A. Della Pietra. 1996. *A Maximum Entropy Approach to Natural Language Processing*. In *Computational Linguistics, 22*.

[Buckwalter2002] Tim Buckwalter. 2002. *Buckwalter Arabic Morphological Analyzer*. In *Linguistic Data Consortium. (LDC2002L49).*

[Chieu and Ng2003] Hai Leong Chieu and Hwee Tou Ng. 2003. *Named Entity Recognition with a Maximum Entropy Approach.* In *Proceedings of CoNLL-2003*.

[Diab et al.2007] Mona Diab, Kadri Hacioglu and Daniel Jurafsky. 2007. *Arabic Computational Morphology: Knowledge-based and Empirical Methods*, chapter 9, pp. 159–179. Abdelhadi Soudi, Antal van den Bosch and Gunter Neumann (Eds.), Springer.

[Diab et al.2004] Mona Diab, Musa Alkhalifa, Sabri Elkateb, Christiane Fellbaum, Aous Mansouri, Martha Palmer. 2007. *Semeval 2007 Task 18: Arabic Semantic Labeling.* In *Proceeding of International Workshop on Semantic Evaluations, SemEval-2007.*

[Farber et al.2006] Benjamin Farber, Dayne Freitag, Nizar Habash and Owen Rambow. 2008. *Improving NER in Arabic Using a Morphological Tagger. Proceedings of the Sixth International Language Resources and Evaluation (LREC'08).*

[Ferrández et al.2007] Sergio Ferrández, Oscar Ferrández, Antonio Ferrández and Rafael Muñoz. 2007. *The Importance of Named Entities in Cross-Lingual Question Answering. In Proc. of Recent Advances in Natural Language Processing, RANLP-2007.*

[Florian et al.2003] Radu Florian, Abe Ittycheriah, Hongyan Jing and Tong Zhang. 2003. *Named Entity Recognition through Classifier Combination. Proc. of CoNLL 2003.*

[Greenwood and Gaizauskas2007] Mark A. Greenwood and Robert Gaizauskas. 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).*

[Habash and Sadat2006] Nizar Habash and Fatiha Sadat. 2006. *Arabic Preprocessing Schemes for Statistical Machine Translation.* In *Proceedings of the North American chapter of the Association for Computational Linguistics (NAACL).*

[Habash and Rambow2005] Nizar Habash and Owen Rambow. 2005. *Arabic Tokenization, Part-Of-Speech Tagging and Morphological Disambiguation in One Fell Swoop*. In *Workshop of Computational Approaches to Semitic Languages, ACL-2005.*

[Klein et al.2004] Dan Klein, Joseph Smarr, Huy Nguyen and Christopher Manning. 2003. *Named Entity Recognition with Character-Level Models. Proc. of CoNLL-2003.*

[Kudo and Matsumato2000] Taku Kudo and Yuji Matsumato. 2000. *Chunking with Support Vector Machine.* In *Proceedings of the 4th Conference on Very Large Corpora*, pages 142-144.

[Kudo et al.2004] Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. 2004. *Applying Conditional Random Fields to Japanese Morphological Analysis*. In *Proceedings of EMNLP*, 2004.

[Lafferty et al.2002] John Lafferty, Andrew McCallum and Fernando Pereira. 2002. *Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data.* In *Proceedings of ICML-2002.*

[Li and McCallum2003] Wei Li and Andrew McCallum. 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.*

[Maamouri et al.2004] Mohamed Maamouri, Ann Bies, Tim Buckwalter and Wig dan Mekki. 2004. *The Penn-Arabic Treebank: Building a large-*

scale annotated Arabic corpus. In *Proceedings of NEMLAR conference on Arabic Language Resourcesand Tools*, 2004.

[Malouf2003] Rob Malouf. 2003. *Markov Models for Language-Independent Named Entity Recognition*. In *Proceedings of CoNLL-2003*. Edmonton, Canada.

[Mayfield et al.2003] James Mayfield, Paul McNamee and Christine Piatko. 2003. *Named Entity Recognition using Hundreds of Thousands of Features*. In *Proceedings of CoNLL-2003*.

[McCallum and Li2003] Andrew McCallum and Wei Li. 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)*.

[Ratnaparkhi1996] Adwait Ratnaparkhi. 1996. *A Maximum Entropy Part-of-Speech Tagger*. In *Proceedings of the First EMNLP*.

[Salton and Buckley1988] Gerard Salton and Chris Buckley. 1988. *Term-weighting Approaches in Automatic Text Retrieval. Information Processing Management*.

[Sha and Pereira2003] Fei Sha and Fernando Pereira. 2003. *Shallow parsing with conditional random fields*. In *Proceedings of HLT-NAACL*.

[Toda and Kataoka2005] Hiroyuki Toda and Ryoji Kataoka. 2005. *A Search Result Clustering Method using Informatively Named Entities.*. In *Proceedings of the 7th ACM International Workshop on Web Information and Data Management*.

[Tran et al.2007] Q. Tri Tran, T.X. Thao Pham, Q. Hung Ngo,Dien Dinh, and Nigel Collier. 2007. *Named Entity Recognition in Vietnamese documents*. *Progress in Informatics Journal*. 2007.

[Thompson and Dozier1997] Paul Thompson and Christopher C. Dozier, 1997. *Name Searching and Information Retrieval. In Proc. of Second Conference on Empirical Methods in Natural Language Processing*,

[Turtle and Croft1991] Howard Turtle and W. Bruce Croft, 1991. *Evaluation of an Inference Network-based Retrieval Model. ACM Transactions on Information Systems*.

[Vapnik1995] Vladimir Vapnik. 1995. *The Nature of Statistical Learning Theory. Springer Verlag*.

[Wu et al.2006] Chia-Wei Wu, Shyh-Yi Jan, Richard Tzong-Han Tsai, and Wen-Lian Hsu. 2006. *On Using Ensemble Methods for Chinese Named Entity Recognition. Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing*. 2006.

[Zhang and Johnson2003] Tong Zhang and David Johnson. 2003. *A Robust Risk Minimization Based Named Entity Recognition System. In: CoNLL-2003*.

[Zitouni et al.2005] Imed Zitouni, Jeff Sorensen, Xiaoqiang Luo and Radu Florian. 2005. *The Impact of Morphological Stemming on Arabic Mention Detection and Coreference Resolution. Proceedings of 43rd Annual Meeting of the Association of Computational Linguistics (ACL05)*. pp. 63-70.

Yassine Benajiba has been granted a scholarship from the Spanish Agency of International Cooperation (AECI) in order to make Ph.D. studies in the Dept. of Informatics and Computation at the Polytechnic University of Valencia under the supervision of Ph.D. Paolo Rosso. He has done some research works in mono-lingual and cross-lingual Question Answering for the Arabic language and he is currently focused on investigating the ”Arabic Named Entity Recognition” task. He has made an internship in CCLS at Columbia University under the supervision of Ph.D. Mona Diab and has been granted a six-month internship in IBM T. J. Watson research center.

Mona Diab received her PhD in 2003 in the Linguistics department and UMIACS, University of Maryland College Park. Her PhD work focused on lexical semantic issues and was titled Word Sense Disambiguation within a Multilingual Framework. Mona is currently an associate research scientist at the Center for Computational Learning Systems, Columbia University. Her research includes work on word sense disambiguation, automatic acquisition of natural language resources such as dictionaries and taxonomies, unsupervised learning methods, lexical semantics, cross language knowledge induction from both parallel and comparable corpora, Arabic NLP in general, tools for processing Arabic(s), computational modeling of Arabic dialects, Arabic syntactic and semantic parsing. She was recently elected to the ACL SIGLEX executive board.

Paolo Rosso received his Ph.D. degree in Computer Science (1999) from the Trinity College Dublin, University of Ireland. He is currently the Head of the Natural Language Engineering Laboratory at the Polytechnic University of Valencia, Spain. His research interests are focused on: word sense disambiguation, clustering narrow domain short texts, plagiarism detection, humour recognition, the Web as lexical resource, multilingual question answering and Arabic Natural Language Processing (NLP). He has published over 90 papers in different conferences, workshops and journals. Paolo Rosso organised several international conferences and workshops also on Arabic NLP (at ICTIS 07 and LREC 08) and he was co-editor of the proceedings of some of these events. He has been involved in a few international research projects with the Arab world: at the moment, he is in charge of a 2nd research project funded by the Spanish Association of International Cooperation AECI between Spain and Morocco on Arabic question answering.

For Review Only

Arabic Named Entity Recognition: A Feature-driven Study

Yassine Benajiba, Mona Diab and Paolo Rosso

For Review Only

Abstract

The Named Entity Recognition (NER) task aims at identifying and classifying Named Entities within an open-domain text. This task has been garnering significant attention recently as it has been shown to help improve the performance of many natural language processing (NLP) applications. In this paper, we investigate the impact of using different sets of features in three discriminative machine learning frameworks, namely, Support Vector Machines, Maximum Entropy and Conditional Random Fields for the task of NER. Our language of interest is Arabic. We explore lexical, contextual and morphological features and nine data-sets of different genres and annotations. We measure the impact of the different features in isolation and incrementally combine them in order to evaluate the robustness to noise of each approach. We achieve the highest performance using a combination of fifteen features in Conditional Random Fields using Broadcast News data ($F_{\beta=1}$=83.34).

Index Terms

Arabic, Natural Language Processing, Named Entity Recognition, Machine Learning Comparison.

I. 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 (NEs) within an open-

Y. Benajiba is a Ph.D. student in the Department of Informatic Systems and Computation at the Polytechnic University of Valencia and a member of the Natural Language Engineering (NLE) Lab. Camino de Vera, s/n, 46022, Valencia, Spain. email: benajibayassine@gmail.com, URL: http://www.dsic.upv.es/∼ybenajiba

M. Diab is an Associate Research Scientist in the Center for Computational Learning Systems at Columbia University. 850 Interchurch Center / MC 7717, 475 Riverside Drive, New York, NY 10115, USA. email: mdiab@ccls.columbia.edu, URL: http://www.cs.columbia.edu/∼mdiab/

P. Rosso is a permanent lecturer in the the Department of Informatic Systems and Computation at the Polytechnic University of Valencia and head of the Natural Language Engineering (NLE) Lab. Camino de Vera, s/n, 46022, Valencia, Spain. email: prosso@dsic.upv.es, URL: http://www.dsic.upv.es/∼prosso

domain text. For instance, for the following text:

'John left Washington D.C. at 4 a.m. and reached New York City at 7h30 a.m.'

A NER system should be able to identify 'John', 'Washington D.C.' and 'New York City' as NEs, and classify the first one as a person and the second and third ones as locations. 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.

We list here some NLP systems that benefit from NER:

- Information Retrieval (IR): The IR system's main goal is to retrieve, from a large document-set, the relevant documents to the user query. In [Thompson and Dozier1997], the authors carried out a statistical study of the percentage of user queries, over a period of several days, to different news databases containing NEs. The authors report that 67.83%, 83.4% and 38.8% of the queries contained a NE in the Wall St. Journal, Los Angeles Times and Washington Post, respectively. In the same paper, the authors report the results obtained when a NER system was integrated in the IR system in order to consider each NE as a single term when the IR system computes the *tf-idf* [Salton and Buckley1988] (term frequency-inverse document frequency). The NE-based approach outperformed the baseline precision (a probabilistic retrieval engine [Turtle and Croft1991]) on all recall levels.
- Question Answering (QA): QA systems aim at answering user specific questions with accurate answers. In [Ferrández et al.2007] the authors argue that a study of the percentage of the questions containing one or more named entities in the CLEF[1] 2004 and 2005 competitions, showed that the majority, precisely 87.7%, of questions contained a NE. Moreover, in [Greenwood and Gaizauskas2007] the authors state that the performance of a QA system can be considerably improved if a NER system is used for a question whose answer is a NE. In their work, the authors show that using an accurate NER system has helped improve the ratio of correctly answered questions of the type *'When did X die?'* from 0% to 53%.
- Machine Translation (MT): The translation of NEs requires different approaches than the translation of common words. For instance, 'Beijing' should be translated as 'Pekin' in French, whereas 'Paris' should remain unchanged. Other NEs need a character-based transliteration such as 'Hizbullah' or

[1] http://www.clef-campaign.org

For Review Only

'Ghandi'. In [Babych and Hartley2003], the authors show that a considerable error rate is experienced when the automatic translator does not manage to transliterate properly the NEs. In order to study the possibility of improving the performance of a MT system by embedding a NER system, they have tagged all the text which has to be translated by a NER system as a pre-processing step. Thereafter, the words tagged by the NER system were translated using the methods which are specific for NEs translation. The results showed that this technique outperforms the methods which do not consider tagging the NEs before the translation.

- Text clustering: Search result clustering (a sub-task of Text Clustering) is the NLP task focused on clustering in groups the results returned by a search engine. For instance, if we have the documents returned by a search engine for the query *'Michael Jordan'*, in order to make these results easier to explore for the user, a result clustering system would cluster the documents concerning Michael Jordan the basketball player[2] in one cluster, and the ones relevant to the Berkeley professor[3] in another cluster. In [Toda and Kataoka2005], the authors report that they have outperformed the existing search results clustering by including a NER system in their global system as it attributes a special weight to the NEs in their clustering approach.

For Review Only

In this paper, we consider the problem of NER for Arabic[4]. The NER task in morphologically rich languages such as Arabic is relatively different from performing the task in English due to inherent characteristic linguistic differences, such as the agglutinative nature of the language allows for complex and hence more sparse data representation. Orthographic characteristics such as lack of capitalization to mark a named entity also render the NER task more challenging in Arabic. We compare between three discriminative approaches to the NER problem: Support Vector Machines (SVMs)[Vapnik1995], Maximum Entropy (ME)[Berger et al.1996] and Conditional Random Fields (CRFs)[Lafferty et al.2002]. We comprehensively investigate many sets of features: contextual, lexical, morphological and shallow syntactic features. We explore the features in isolation first. Thereafter we rank the features according to their impact and we evaluate the approaches using the N features with the highest impact. We have conducted experiments for N=*1* to N=*total number of features* in order to show the robustness to noise of each of the mentioned approaches. We experiment with two sets of data, the standard ACE data and a manually created data set, NLE-corpus, in order to confirm the reliability of our results. Our best system

[2] http://en.wikipedia.org/wiki/Michael_Jordan

[3] http://www.cs.berkeley.edu/∼jordan/

[4] We use Arabic in this paper to refer to Modern Standard Arabic.

that combines the fifteen best features using CRFs yields an overall F1 score of 83.34.

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

II. Related Work

For Review Only

There are several significant research efforts in NER. In the Conference on Natural Language Learning (CoNLL) 2003 NER evaluation tasks[5] the participants were asked to identify the NEs within the test-sets (English and German) and classify them. The system which has yielded the best performance is described in [Florian et al.2003]. The authors report that they have used a linear interpolation of three different classifiers: (i) Hidden Markov Models; (ii) Maximum Entropy (ME); and (iii) Robust Risk Minimization (RRM). Their final results were 88.76 for English and 72.41 for German, best results in both languages. [Chieu and Ng2003] was ranked the second best participation. The system is fully ME-based and uses different types of features, namely, contextual and lexical features as well as capitalization. The authors performed two runs where the second one uses additional external resources. For English, the results using the lexicon (88.12) were almost two points higher than those obtained without using an external resource (86.83). However, the results were higher when no external resource was used for German (77.05 vs. 76.83). Finally, the system of [Klein et al.2004] was ranked third. It employed a character-based HMM approach. This method relies heavily on internal evidence for the NEs. Their final results for English were 86.07 (third best results) and for German 71.9 (second best results).

[Tran et al.2007] show that using a Support Vector Machine (SVM) approach outperforms ($F_{\beta=1}$=87.75) using CRFs (86.48) on the NER task in Vietnamese. However, this comparison is based on the average F-measure obtained by using the same feature-set with both SVMs and CRFs. Thus it is not possible to make any further conclusions on the behavior of these Machine Learning (ML) approaches with different features. In [Zhang and Johnson2003], the authors report a manual feature selection study in which they prove that simple token-based features can be more helpful than sohpisticated linguistic features. In this study, we can find the performance obtained with different feature-sets, however the authors did not

[5] http://www.cnts.ua.ac.be/conll2003

compare the performance of the different ML approaches.

With current surge in resources making their way in the NLP community for Arabic, we are starting to see systems being developed for the processing of the Arabic language.

In work by [Zitouni et al.2005], the authors use a 2-step approach (NE boundary detection and then NE classification) in their investigation of Arabic mention detection problem. A mention refers to a named entity (e.g. Ohio), a nominal (e.g. Prime Minister), or a pronominal (e.g. he) entity. They pre-process the data applying morphological stemming. They adopt an ME Markov model approach exploring lexical, syntactic and gazetteer features. The authors evaluate their system's performance against the Automatic Content Extraction (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, i.e. we are not able to tell the performance on NER task specifically. On the task of Arabic NER alone, there has been recent significant work. In some of our recent work, [Benajiba et al.2007], we show that using a basic ME approach to Arabic NER yields an F1-measure of 55.23. We evaluate the system using a corpus that was developed in-house using CoNLL guidelines, NLE-corpus. We followed up with further work in [Benajiba and Rosso2007] which yields results reaching $F_{\beta=1}$=65.91 by adopting a two stage classification using an ME based approach to the problem in a style similar to [Zitouni et al.2005]. Finally, in our most recent work on the problem, we explore using CRFs in [Benajiba and Rosso2008] with different features, namely, contextual, morphological, and lexical features, together with gazetteers as external resources. We report the performance obtained with each feature in isolation and our best results were acheived when all the features were combined (79.21). Finally, [Farber et al.2006] use a structured perceptron as well as sophisticated morphological features for the task of Arabic NER. The authors report achieving an F-measure of 75.7 on the newswire subset of the ACE 2005 corpora.

For Review Only

III. Arabic in the context of Named Entity Recognition Task

Let us consider the example presented in Figure 1. In Buckwalter transliteration it can be written as *whw llsnp bmvAbp Alqmr ynyr lylA lyzyn AlsmA'*[6]. The underlined word ('llsnp') can be read as 'to the year' or 'to the Sunnah', i.e. in the latter case it is an NE. If we consider that the first case to be correct then the whole sentence can be translated as 'and it is to the year as the moon which lightens in the night to beautify the sky', and 'it' could refer to a special month or day in the year. In the second case,

[6] For purposes of presentation, we adopt a Buckwalter transliteration scheme to show romanized Arabic [Buckwalter2002].

وهو للّسنة بمثابة القمر ينير ليلا ليزين السماء

Fig. 1. An illustrating example of the difficulty of Arabic NER

For Review Only

the correct translation would be ‘and he is to the Sunnah as the moon which lightens in the night to beautify the sky’, and ‘he’ could refer to someone of great importance for the ‘Sunnis’. Hence, if we do not have any other information it would not be possible to disambiguate whether ‘llsnp’ is a NE or not for the two following reasons:

- *Absence of short vowels:* In newpaper articles, magazines, books and all resources written in MSA, the texts are mostly unvocalized. The vocalized form of the word ‘llsnp’ is spelled differently, ‘llsunnap’ (meaning ‘for the Sunnah’) vs. ‘llsanap’ (meaning ‘for the year’) disambiguating between the two readings. Human readers use context and knowledge in order to disambiguate the unvocalized forms. However, in the case of an NER system, the considered context is typically limited and the knowledge (lexicons, gazetteers, etc.) sources are inherently static and limited in scope.
- *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. Hence, in our example there is no way we can capitalize (or mark) the NE given the Arabic orthography rules.
- *Sparseness:*From a statistical NLP viewpoint, morphologically rich languages have another more important obstacle generally known as ‘data sparseness’. These languages use an agglutinative strategy to form surface tokens. In the case of Arabic, being a Semitic language,[7] it exhibits a templatic morphology where words are made up of roots and affixes. Clitics agglutinate to words. For instance, the surface word in Figure 2 *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’.

وبحسناتهم

Fig. 2. An illustrating example of the templatic morphology of the Arabic language

[7] Other Semitic languages include Hebrew and Amharic

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 (from a statistical viewpoint) is much more sparse compared to English. In order to tackle this problem, we need to perform a level of *segmentation of the clitics for each word* (*clitic-segmentation*) as a pre-processing step. It is particularly useful for NER as: (i) the NEs always appear in the same form hence lowering the number of unseen NEs; (ii) it reduces the number of different surface form contexts in which the NEs appear.

IV. Approach using A Large Range of Features

A. *The SVMs, ME and CRFs Approach*

For Review Only

In this paper, we explore three different yet comparable approaches that were used for NER in other languages: SVMs, ME, and CRFs. The approaches have well known desirable characteristics for NLP applications.

SVMs are proven to be robust to noise and to have a 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 [Diab et al.2004], [Kudo and Matsumato2000], and for the NER task in particular [Mayfield et al.2003], [Tran et al.2007]. In order to use SVMs in the NER task, we use *Yamcha*[8] toolkit.

ME aims at providing a model with the less biases possible [Berger et al.1996]. One of the first implementations of the ME approach in NLP tasks is [Ratnaparkhi1996]. Moreover, as mentioned in section II, this approach proved to be successful for the NER task. We implemented our own ME approach to carry out the experiments, for weight estimation we use *Yasmet*.[9]

CRFs are oriented to segmenting and labeling sequence data [Lafferty et al.2002]. As undirected graphical models, CRFs are alternative ways to represent probability distributions. During the training phase the conditional probabilities of the classes are maximized. CRFs are proven to be very efficient for the NER task (see section II). We use *CRF++*[10] for our experiments.

B. *Arabic NER Task Tag sets*

While different NLP applications may require different tag sets, we address here the NER task as an end system in itself. There exist three standard NER tag sets in the literature:

[8] http://chasen.org/∼taku/software/yamcha/

[9] http://www.fjoch.com/YASMET.html

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

(i) Message Understanding Conference (MUC-6)[11]: the NER task consisted of three subtasks:

- ENAMEX: for proper nouns;
- NUMEX : for numerical expressions; and
- TIMEX : for temporal expressions.

The ENAMEX subtask was defined as the identification of the NEs and their classification:

- Person, e.g. Albert Einstein;
- Location, e.g. Paris;
- Organization, e.g. Google Co.

(ii) Conference of Natural Language Learning (CoNLL): In the language-independent NER shared task held in the CoNLL 2002[12] and CoNLL 2003[13] the tag-set comprised four classes: Person, Location, Organization (same as previous ones) and *Miscellaneous (e.g. Empire State building)*;

(iii) ACE: The ACE 2003 data defines four different classes: Person, Geographical and Political Entities (GPE), Organization and Facility. Whereas in ACE 2004 and 2005 two classes were added to the previous 2003 tag set: *Vehicles (e.g. Rotterdam Ship)* and *Weapons (e.g. Kalashnikof)*.

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.

C. Features

The most challenging aspect of any machine learning approach to NLP problems is deciding on the optimal feature sets. In this work, we investigate a large space of features. The feature sets are 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_i):** This feature defines the lexical orthographic nature of the tokens in the text. We define it as a character n-gram of 6 characters. This feature is elaborated in the following

[11] http://cs.nyu.edu/cs/faculty/grishman/muc6.html

[12] http://www.cnts.ua.ac.be/conll2002/

[13] http://www.cnts.ua.ac.be/conll2003/

For Review Only

example. Consider that a word is simply a sequence of characters $C_1C_2C_3...C_{n-1}Cn$ then the lexical features would be

- LEX_1=C_1
- LEX_2=C_1C_2
- LEX_3=$C_1C_2C_3$
- $LEX4$=C_n
- $LEX_5 = C_{n-1}C_n$
- $LEX_6 = C_{n-2}C_{n-1}C_n$

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[14] as well as other web sources. This enrichment consisted of: (i) taking the page labeled '*Countries of the world*' (dwl AlEAlm) as a starting point to crawl into Wikipedia and retrieve location names; (ii) we automatically filter the data removing stop words; (iii) finally, we manually filter the resulting set ensuring its good quality as a source of location names.

For Review Only

Morphological features (M_x): This feature set is based on exploiting the rich characteristic morphological features of the Arabic language. We relied on a system for Morphological Analysis and Disambiguation for Arabic (MADA) to extract relevant morphological information [Habash and Rambow2005]. 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 raw 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-segmentation.

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. Although MADA produces fourteen features we only use eleven of them, the description of each one of these features goes beyond the scope of this paper, yet they are explained in detail in [Habash and Rambow2005]. However, the eleven features used in our experiments are listed as follows:

- M_{ART}=article: indicates whether a token has a definite article or not;

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

- M_{ASP}= verb aspect: In Arabic, a verb maybe imperfective, perfective or imperative.
- M_{CASE}=grammatical case: genitive, accusative, nominative;
- M_{CLIT}=clitic: indicates whether a word has any clitics attached;
- M_{CONJ}=conjunction: MADA indicates whether a token has any conjunction attached or not;
- M_{DEF}=definiteness: MADA indicates whether a token is definite or not. All the NEs by definition are definite;
- M_{MOOD}=mood: indicative, imperative, subjuntive and optative are the possible values of this feature;
- M_{NUM}=number: For almost all the tokens categories (verbs, nouns, adjectives, etc.) MADA provides the grammatical *number*. In Arabic, the possible values are singular (SG), dual (DU) and plural (PL);
- M_{PART}=particle: MADA indicates whether a token has any particles attached or not;
- M_{PER}=person: In Arabic, verbs, nouns, and pronouns typically indicate person information. The possible values are *first, second* or *third* person;
- M_{VCE}=voice: In Arabic, a verb can have one of two possible voice values: active or passive.

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[15] described in [Diab et al.2007]. Like the MADA system, AMIRA-1.0 is an SVM based set of tools. The POS tagger performs at 96.2% and the BPC system performs at 96.33%. 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 POS tag set of 25 tags created for parsing and included in the Arabic Treebank [Maamouri et al.2004] distribution.

For Review Only

Nationality (NAT): We mark nationalities in the input text. Such information is useful for detecting NEs since they are used as precursors to recognize NE. Especially, when location and person NEs are introduced in a text they are usually preceeded by a nationality. For instance, الرئيس **الامريكي** $\overline{\text{جورج بوش}}$, which can be trasliterated as “Alr¿ys **AlAmryky** *jwrj bw$*”, and translated to English as “The **American** President *George Bush*”. Another example is العاصمة **الاسبانية** $\overline{\text{مدريد}}$, which can be transliterated as “AlEAsmp **AlASbanyp** *mdryd*” and translated to English as “the **Spanish** capital *Madrid*”.

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

[15] http://www1.cs.columbia.edu/∼mdiab/software/AMIRA-1.0.tar.gz

process it taps into an underlying lexicon that provides bilingual information. The insight is that if the translation begins with a capital letter, then it is most probably a NE.

V. Experiments and Results

A. Data

We use the ACE 2003, 2004 and 2005 corpora and an enhanced version of the corpus used in [Benajiba et al.2007], NLE-corpus. Table I describe for the different corpora: the training data size ($Size_{train}$), the test size ($Size_{test}$).

NLE-corpus

For Review Only

This corpus (see Table I) 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 CoNLL tag set as described in Section IV-B. The CoNLL tag set comprises 4 classes: Person, Location, Organization and Miscellaneous. The annotators followed the IOB2 annotation guidelines [Ratnaparkhi1996]. A NE can comprise more than one token. The IOB2 annotation scheme tags the first token in a NE as $B - Class$, $I - Class$ for each subsequent token, and O for the tokens which are not part of any NE. For instance:

'John left Washington D.C. at 4 a.m. and reached New York City at 7h30 a.m.'

the IOB2 annotation scheme tags it as:

'John/**B-PER** left Washington/**B-LOC** D.C./**I-LOC** at 4 a.m. and reached New/**B-LOC** York/**I-LOC** City/**I-LOC** at 7h30 a.m.'

ACE data

The ACE data (see Table I) is annotated for many tasks: Entity Detection and Tracking (EDT), Relation Detection and Recognition (RDR), Event Detection and Recognition (EDR). The ACE 2003 comprises two data genres: *Broadcast News* (BN) and *Newswire* (NW). The ACE 2004 additionally comprises the *Arabic Treebank* (ATB). The ACE 2005 does not have the ATB genre but it has *Weblogs* (WL). The ACE data annotates pronominal, nominal and named mentions of entities, since it caters to more than one type of task. For our purposes, we specifically care about the NER aspect of the data, hence, we discard the annotations of pronominal and nominal mentions and keep only the named mentions.

The only main difference which remains between the ACE annotation and the NLE-corpus ones is that

$Corpus$	$genre$	$Size_{train}$	$Size_{test}$
NLE-corpus	NW	144.48k	30.28k
ACE 2003	BN	16.34k	2.51k
	NW	29.44k	7k
ACE 2004	BN	50.44k	13.32k
	NW	51.74k	13.4k
	ATB	21.27k	5.25k
ACE 2005	BN	22.3k	5k
	NW	43.85k	12.3k
	WL	18k	3.2k

TABLE I

Characteristics of NLE-Corpus and ACE 2003, 2004 and 2005 data

For Review Only

in the former, nationalities (e.g. Spanish, French, etc.) are tagged as geopolitical entities (GPE) whereas in the ACE annotation scheme they are not considered NEs.

B. Experimental Set-up

1) Metrics: We use the CoNLL[16] evaluation standard metrics of precision, recall and F1-measure, which is the harmonic mean between precision and recall.

The CoNLL evaluation metrics are aggressive metrics in that they do not assign partial credit. A NE has to be identified as a whole (full span) and correctly classified in order to obtain credit.

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

a- 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 (see Table IV) indicate the percentage of NEs already seen in the training phase.

b- Parameter setting:

We need to establish the impact of two experimental factors on NER performance, namely clitic-segmentation and contextual window size as a preliminary pre-cursor to our feature engineering experiments. Clitic-segmentation in a highly agglutinative language such as Arabic has been shown to be useful

[16] http://www.cnts.ua.ac.be/conll2003/

for many NLP applications [Habash and Sadat2006]. Intuitively, clitic-segmentation serves as a first layer of smoothing in such sparse high dimensional spaces. We need to decide on an optimal window size, so we experiment with different sizes. We set the clitic-segmentation to the ATB standard clitic-segmentation scheme.[17] In these experiments, we investigate window sizes of $-1/+1$ up to $-4/+4$ tokens/words surrounding a target NE. We carry out the experiments on the NLE-corpus using SVMs without any additional features. Table II shows the CoNLL results obtained for the raw text corpus (UNSEG) and the segmented clitics corpus (SEG), respectively.

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

TABLE II

PARAMETER SETTING EXPERIMENTS: COMPARISON BETWEEN DIFFERENT WINDOW SIZES, AND THE IMPACT OF CLITIC-SEGMENTATION ON THE NER TASK

For Review Only

From Table II we note that clitic-segmentation has a significant positive impact on NER. We see an increase of 3 absolute points in F1 score when the text is clitic segmented. 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.

c- Feature engineering:

We conduct different sets of experiments to explore the space of possible features. We use clitic segmented text and we define the context (CXT) to be $-1/+1$ as established in the previous section. The rest of our experiments are organized as follows:

1) Explore individual features[18]: which consists of measuring the impact (number of F-measure points of improvement obtained) when each of the feature is used separately using each of the ML approaches and data-sets which we have described earlier;
2) Rank features according to their impact: from the obtained results in the previous step, we get a ranking of the features for each ML approach and data-set. In this step we aim at deducing a

[17] The ATB clitic-segmentation scheme consists of separating from the stem word: (i) prefixes are typically conjunctions and preposition; and (ii) pronominal suffixes.

[18] Due to space limitations, the results are not presented in this paper.

general ranking of the features. The algorithm which we have used for this purpose, assigns to each feature the most frequent rank. Table III illustrates the ranking obtained in our experiments.

3) Evaluate SVMs, ME and CRFs approaches combining each time the top N-top elements of the ranked features list. We have carried out experiments starting from N=*1* and up to from N=*22* to find the optimal number of features.

Rank	Feature	Rank	Feature
1	POS	12	NAT
2	CAP	13	LEX_1
3	M_{ASP}	14	LEX_4
4	M_{PART}	15	M_{CASE}
5	LEX_6	16	M_{NUM}
6	LEX_3	17	M_{DEF}
7	M_{CLIT}	18	LEX_2
8	BPC	19	LEX_5
9	GAZ	20	M_{CONJ}
10	M_{ART}	21	M_{MOOD}
11	M_{VCE}	22	M_{PER}

TABLE III

FEATURES RANKED ACCORDING TO THEIR IMPACT.

For Review Only

We evaluate the performance in our experiments using 5-fold cross validation on each corpus independently. For the NLE-corpus we have chosen the same ratio of test data size to training data size which has been used in the CoNLL competitions. For the ACE data, we have replicated the same splits which were adopted in the ACE evaluations. (Table I shows the average size of the training and test data for each corpus).

Figure 3 shows the results for ACE 2003 data, BN genre (best results). Figure 4 illustrates the results for the ACE 2004 data. Figure 5 shows the results obtained with ACE 2005, WL genre (worst results). Finally, table IV presents the baseline and the best results obtained for each corpus together with the number of features N and the corresponding ML approach. In the same table we also present the results which were obtained when all the features were combined.

Corpus	**genre**	**Baseline**	Best						All Features		
			SVMs		ME		CRFs		SVMs	ME	CRFs
			N	F-score	N	F-score	N	F-score			
NLE-corpus	NW	31.5	14	**81.04**	3	77.9	12	80.36	80.4	76.8	79.8
ACE 2003	BN	74.78	15	82.72	3	78.05	15	**83.34**	82.71	74.84	82.94
	NW	69.08	14	**79.72**	3	74.56	13	79.52	79.21	73.84	79.11
ACE 2004	BN	62.02	16	**77.61**	2	73.34	13	77.03	76.43	69.44	76.96
	NW	52.23	14	74.13	3	68.13	12	**74.53**	73.4	63.13	73.47
	ATB	64.23	15	75.43	2	69.95	13	**75.51**	75.34	64.66	75.48
ACE 2005	BN	71.06	15	**82.02**	3	77.67	14	81.87	81.47	75.71	81.1
	NW	58.63	15	76.97	3	70.31	13	**77.06**	76.19	67.41	75.67
	WL	27.66	12	**55.69**	2	44.96	14	53.91	53.81	32.66	51.81

TABLE IV

BEST OBTAINED RESULTS FOR EACH CORPUS.

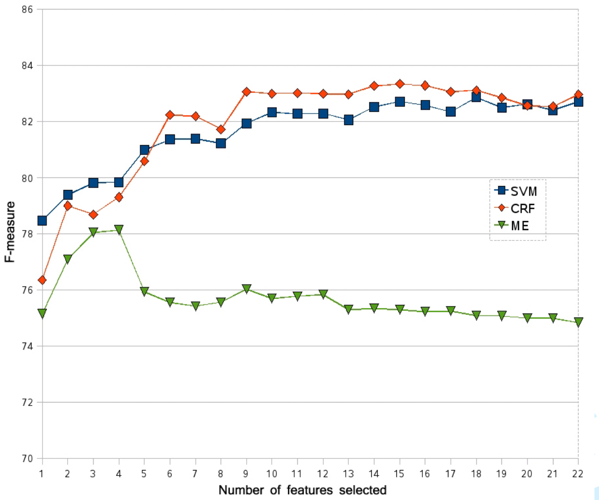

Fig. 3. Results per approach and number of features for the ACE 2003 (Broadcast News genre) data.

VI. DISCUSSION AND ERROR ANALYSIS

We achieve state-of-art and significantly improve over the baseline for almost all the corpora. We have obtained an F1 score up to 83.34 for ACE 2003, Broadcast News genre. The worst results are yielded for the WL genre of data which may be explained by the overall randomness of the WL data

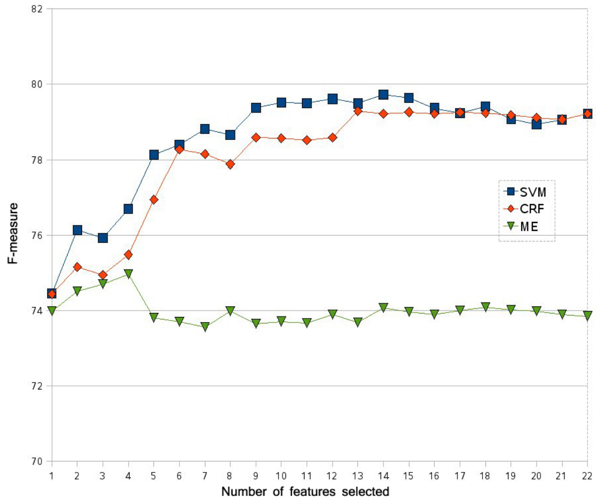

Fig. 4. Results per approach and number of features for the ACE 2003 (Newswire genre) data.

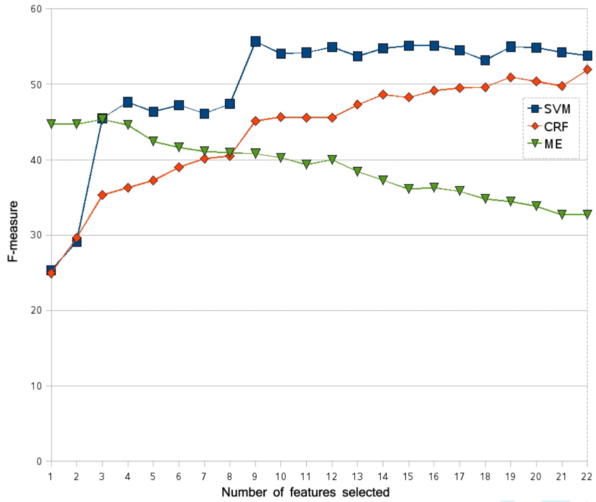

Fig. 5. Results per approach and number of features for the ACE 2005 (Weblogs genre) data.

relative to the other genres, in addition to the fact that WL data includes dialectal language which plays havoc with the basic processing tools such POS tagging and morphological disambiguation.Farber et al., [Farber et al.2006], report an F-measure of 75.7 on the NW genre of the ACE 2005 data using a set of morphological features. Our approach outperforms theirs, yielding an F-measure of 77.06 (i.e. 1.36 points absolute difference) for the same data-set by using CRFs with the 13 best features (see Table III).

Features:

Let us consider the following sentence:

الرئيس الروسي فلاديمير ب # وطن في اقرب ...

in Buckwalter transliteration as:

Alrys Alrwsy flAdymyr b# wTn fy Aqrb ...

tranlated to English as:

The Russian president Vladimir Putin in the nearest ...

For Review Only

The word 'Putin', which in Arabic is generally spelled بوتن (*bwtn*) is spelled differently in that specific text as بوطن(*bwTn*). Moreover, the clitic segmenter mistakenly split the first character from the word treating it as a prepositional prefix ب, meaning 'in' or 'with'. Furthermore, the wrong spelling of *wTn* confuses the system further with the word meaning country which has the same spelling. Hence the word (*bwTn*) is misclassified as an O. Even if a NE can accept a prefix, it attaches to the first token in the NE.

The CAP feature is ranked second (see table III) among all others. This result confirms that the lack of capitalization in some languages such as Arabic considerably complicates the NER task. The use of lexical features (LEX_i) shows that only marking the first and last three characters of a word (LEX_3 and LEX_6) can be useful for a NER approach. The rest of the lexical features occur randomly with all the classes. The lexical features are mostly useful when the same NE appears slightly different in the different parts of the corpus. For instance, in the sentence:

تقدم ايريل شارون ...

transliterated as:

tqdm Ayryl $Arwn ...

translated to English as:

Ariel Sharon presented ...

The name 'Ariel' is transliterated in Arabic as اريل (*Aryl*) or ايريل (*Ayryl*). In the training corpus, this name appears only with the first transliteration. Hence, when seen in the latter spelling form in the test data, the classifier assigns it an O tag. However, when we employ the last trigram of the word as a feature, (LEX_6), the classifier correctly classifies the alternate spelling, ايريل (*Ayryl*) as part of a NE. It shares the same last three characters with the word اريل (*Aryl*) which has been frequently seen as a person in the training data. Another similar example is the NE الواشنطن بوست (*AlwA$ntn bwst*) 'The Washington Post' which appears with the definite article (*Al*) only once in the corpus. The classifier tags the word correctly only when the lexical feature LEX_6 is used.

On the other hand, the lexical feature LEX_3, which concerns the first three characters of each word, is mostly useful for NEs with different suffixes. The most remarkable example of such NEs are the nationalities which are tagged (depending on the context) as LOC, PER or O. Similar to English the difference between plural and singular forms of most of the nationalities is the suffix (e.g. 'Palestinian', فليسطيني (*flsTyny*) vs. 'Palestinian**s**', فليسطيني ين (*flsTyny***yn**). In addition, the Arabic has the dual form which is very rarely used but also requires only adding a suffix to the singular form. Hence ignoring the suffixes and focusing on first 3 letters in a token works in our favor in the case of nationalities. In our data, LEX_i features are very useful in capturing those cases.

For Review Only

Incremental Features Selection: The incremental feature selection yields slightly better results than using all features together. Moreover, it is important to notice that the time to extract, train and test with only 14 or 15 features is almost half the time necessary for 22 features. Through the examples of errors corrected when the best feature-set is used, we note that simply when we use a selected feature-set, we avoid providing the classifier noisy information. One case is the NE 'Holy Shrine', a facility FAC, in Arabic الحرم القدسي (*AlHrm Alqdsy*) is correctly classified with the best feature set. If all the features were to be used, for example the starting bigram, the definite article 'Al', the classifier mis-tags the tokens as O.

Approaches: The results obtained with ME are considerably lower than the ones obtained by CRFs and SVMs especially when the number of features exceed 6. This shows that the ME approach is much more sensitive to noise and that it is more suitable to use this approach when a restricted number of accurate features is used. On the other hand, CRFs and SVMs show very similar performance. Even though SVMs show a slightly better performance when only the first 7 top features are used. Thus, according to our results it is not possible to determine an absolute superiority of the SVMs or the CRFs for the Arabic NER task. Through the data we have also observed that even if SVMs and CRFs give different 'false alarms' they tend to miss the same NEs. Hence there is no real complementarity between the approaches.

Accordingly, the choice of one or the other has to be based on the number of available features and their quality.

VII. Conclusions and Future Directions

We describe the performance obtained using language-dependent and language independent features in SVMs, ME and CRFs for the NER task on different Arabic data-sets of different genres. We measure the impact of each feature individually, we rank them according to their impact and then perform incremental features' selection considering each time the N best features (we exhaustively explore all the possible values of N). Our experiments yield state of the art performance significantly outperforming the baseline. Our best results achieve an F1 score of 83.34 using the 15 best features in the CRFs approach on the ACE 2003 BN data. Our results show that the SVMs and CRFs have very similar behaviors and significantly outperform the ME approach. They also strongly suggest that the choice of using the CRFs or the SVMs approach should be based on the number of features available, i.e. if only few features are avaailable it is more suitable to use SVMs. Using the Arabic language in our experiments showed that a better performance is obtained in the NER task for languages which exhibit complex and rich structures if the data is pre-processed by a clitic-segmenter. They also show these languages can profit from their morphological richness if morphological features for each word are extracted and provided to the classifier. Those morphological features help the classifier by indicating that a word has a certain number of characteristics which makes it more or less probable to be a NE.

Other features such as the lexical features, which are based on the starting and ending character trigrams of the word, are totally language-independent, very easy to extract and they show that they can be very useful to capture NEs which might appear with a slight difference in the surface form in their occurrences in the data.

For Review Only

Acknowledgments

Acknowledgments here

References

[Babych and Hartley2003] Bogdan Babych and Anthony Hartley. 2003. *Improving Machine Translation Quality with Automatic Named Entity Recognition*. In *Proc. of EACL-EAMT*. Budapest.

[Benajiba and Rosso2008] Yassine Benajiba and Paolo Rosso. 2008. *Arabic Named Entity Recognition using Conditional Random Fields*. In *In: Proc. Workshop on HLT NLP within the Arabic world. Arabic Language and local languages processing: Status Updates and Prospects, 6th Int. Conf. on Language Resources and Evaluation, LREC-2008.*

[Benajiba et al.2007] Yassine Benajiba, Paolo Rosso and José Miguel Benedí. 2007. *ANERsys: An Arabic Named Entity Recognition system based on Maximum Entropy*. In *Proceedings of 8th Int. Conf. on Comput. Linguistics and Intelligent Text Processing, CICLing-2007*, Springer-Verlag, LNCS(4394), pp. 143-153.

[Benajiba and Rosso2007] Yassine Benajiba and Paolo Rosso. 2007. *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, 3rd Indian Int. Conf. on Artificial Intelligence, IICAI-2007*.

[Bender et al.2003] Oliver Bender, Franz Josef Och, and Hermann Ney. 2003. *Maximum Entropy Models For Named Entity Recognition*. In *Proceedings of CoNLL-2003*. Edmonton, Canada.

[Berger et al.1996] Adam L. Berger, Vincent J. Della Pietra and Stephen A. Della Pietra. 1996. *A Maximum Entropy Approach to Natural Language Processing*. In *Computational Linguistics, 22*.

[Buckwalter2002] Tim Buckwalter. 2002. *Buckwalter Arabic Morphological Analyzer*. In *Linguistic Data Consortium. (LDC2002L49)*.

[Chieu and Ng2003] Hai Leong Chieu and Hwee Tou Ng. 2003. *Named Entity Recognition with a Maximum Entropy Approach*. In *Proceedings of CoNLL-2003*. Edmonton, Canada.

[Diab et al.2007] Mona Diab, Kadri Hacioglu and Daniel Jurafsky. 2007. *Arabic Computational Morphology: Knowledge-based and Empirical Methods*, chapter 9, pp. 159–179. Abdelhadi Soudi, Antal van den Bosch and Gunter Neumann (Eds.), Springer.

[Diab et al.2004] Mona Diab, Musa Alkhalifa, Sabri Elkateb, Christiane Fellbaum, Aous Mansouri, Martha Palmer. 2007. *Semeval 2007 Task 18: Arabic Semantic Labeling*. In *Proceeding of International Workshop on Semantic Evaluations, SemEval-2007*.

[Farber et al.2006] Benjamin Farber, Dayne Freitag, Nizar Habash and Owen Rambow. 2008. *Improving NER in Arabic Using a Morphological Tagger. Proceedings of the Sixth International Language Resources and Evaluation (LREC'08).*

[Ferrández et al.2007] Sergio Ferrández, Oscar Ferrández, Antonio Ferrández and Rafael Muñoz. 2007. *The Importance of Named Entities in Cross-Lingual Question Answering. In Proc. of Recent Advances in Natural Language Processing, RANLP-2007.*

[Florian et al.2003] Radu Florian, Abe Ittycheriah, Hongyan Jing and Tong Zhang. 2003. *Named Entity Recognition through Classifier Combination. Proc. of CoNLL 2003.*

[Greenwood and Gaizauskas2007] Mark A. Greenwood and Robert Gaizauskas. 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).*

[Habash and Sadat2006] Nizar Habash and Fatiha Sadat. 2006. *Arabic Preprocessing Schemes for Statistical Machine Translation*. In *Proceedings of the North American chapter of the Association for Computational Linguistics (NAACL).*

[Habash and Rambow2005] Nizar Habash and Owen Rambow. 2005. *Arabic Tokenization, Part-Of-Speech Tagging and Morphological Disambiguation in One Fell Swoop*. In *Workshop of Computational Approaches to Semitic Languages, ACL-2005*.

[Klein et al.2004] Dan Klein, Joseph Smarr, Huy Nguyen and Christopher Manning. 2003. *Named Entity Recognition with Character-Level Models. Proc. of CoNLL-2003.*

[Kudo and Matsumato2000] Taku Kudo and Yuji Matsumato. 2000. *Chunking with Support Vector Machine*. In *Proceedings of the 4th Conference on Very Large Corpora*, pages 142-144.

For Review Only

[Kudo et al.2004] Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. 2004. *Applying Conditional Random Fields to Japanese Morphological Analysis*. In *Proceedings of EMNLP*, 2004.

[Lafferty et al.2002] John Lafferty, Andrew McCallum and Fernando Pereira. 2002. *Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data*. In *Proceedings of ICML-2002*.

[Li and McCallum2003] Wei Li and Andrew McCallum. 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*.

[Maamouri et al.2004] Mohamed Maamouri, Ann Bies, Tim Buckwalter and Wig dan Mekki. 2004. *The Penn-Arabic Treebank: Building a large-scale annotated Arabic corpus*. In *Proceedings of NEMLAR conference on Arabic Language Resourcesand Tools*, 2004.

[Malouf2003] Rob Malouf. 2003. *Markov Models for Language-Independent Named Entity Recognition*. In *Proceedings of CoNLL-2003*. Edmonton, Canada.

[Mayfield et al.2003] James Mayfield, Paul McNamee and Christine Piatko. 2003. *Named Entity Recognition using Hundreds of Thousands of Features*. In *Proceedings of CoNLL-2003*.

[McCallum and Li2003] Andrew McCallum and Wei Li. 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)*.

[Ratnaparkhi1996] Adwait Ratnaparkhi. 1996. *A Maximum Entropy Part-of-Speech Tagger*. In *Proceedings of the First EMNLP*.

[Salton and Buckley1988] Gerard Salton and Chris Buckley. 1988. *Term-weighting Approaches in Automatic Text Retrieval. Information Processing Management*.

[Sha and Pereira2003] Fei Sha and Fernando Pereira. 2003. *Shallow parsing with conditional random fields*. In *Proceedings of HLT-NAACL*.

[Toda and Kataoka2005] Hiroyuki Toda and Ryoji Kataoka. 2005. *A Search Result Clustering Method using Informatively Named Entities.*. In *Proceedings of the 7th ACM International Workshop on Web Information and Data Management*.

[Tran et al.2007] Q. Tri Tran, T.X. Thao Pham, Q. Hung Ngo,Dien Dinh, and Nigel Collier. 2007. *Named Entity Recognition in Vietnamese documents*. *Progress in Informatics Journal*. 2007.

[Thompson and Dozier1997] Paul Thompson and Christopher C. Dozier, 1997. *Name Searching and Information Retrieval. In Proc. of Second Conference on Empirical Methods in Natural Language Processing*,

[Turtle and Croft1991] Howard Turtle and W. Bruce Croft, 1991. *Evaluation of an Inference Network-based Retrieval Model. ACM Transactions on Information Systems*.

[Vapnik1995] Vladimir Vapnik. 1995. *The Nature of Statistical Learning Theory. Springer Verlag*.

[Wu et al.2006] Chia-Wei Wu, Shyh-Yi Jan, Richard Tzong-Han Tsai, and Wen-Lian Hsu. 2006. *On Using Ensemble Methods for Chinese Named Entity Recognition. Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing*. 2006.

[Zhang and Johnson2003] Tong Zhang and David Johnson. 2003. *A Robust Risk Minimization Based Named Entity Recognition System. In: CoNLL-2003*.

[Zitouni et al.2005] Imed Zitouni, Jeff Sorensen, Xiaoqiang Luo and Radu Florian. 2005. *The Impact of Morphological Stemming on Arabic Mention Detection and Coreference Resolution. Proceedings of 43rd Annual Meeting of the Association of Computational Linguistics (ACL05)*. pp. 63-70.

For Review Only

Yassine Benajiba has been granted a scholarship from the Spanish Agency of International Cooperation (AECI) in order to make Ph.D. studies in the Dept. of Informatics and Computation at the Polytechnic University of Valencia under the supervision of Ph.D. Paolo Rosso. He has done some research works in mono-lingual and cross-lingual Question Answering for the Arabic language and he is currently focused on investigating the ”Arabic Named Entity Recognition” task. He has made an internship in CCLS at Columbia University under the supervision of Ph.D. Mona Diab and has been granted a six-month internship in IBM T. J. Watson research center.

For Review Only

Mona Diab received her PhD in 2003 in the Linguistics department and UMIACS, University of Maryland College Park. Her PhD work focused on lexical semantic issues and was titled Word Sense Disambiguation within a Multilingual Framework. Mona is currently an associate research scientist at the Center for Computational Learning Systems, Columbia University. Her research includes work on word sense disambiguation, automatic acquisition of natural language resources such as dictionaries and taxonomies, unsupervised learning methods, lexical semantics, cross language knowledge induction from both parallel and comparable corpora, Arabic NLP in general, tools for processing Arabic(s), computational modeling of Arabic dialects, Arabic syntactic and semantic parsing. She was recently elected to the ACL SIGLEX executive board.

Paolo Rosso received his Ph.D. degree in Computer Science (1999) from the Trinity College Dublin, University of Ireland. He is currently the Head of the Natural Language Engineering Laboratory at the Polytechnic University of Valencia, Spain. His research interests are focused on: word sense disambiguation, clustering narrow domain short texts, plagiarism detection, humour recognition, the Web as lexical resource, multilingual question answering and Arabic Natural Language Processing (NLP). He has published over 90 papers in different conferences, workshops and journals. Paolo Rosso organised several international conferences and workshops also on Arabic NLP (at ICTIS 07 and LREC 08) and he was co-editor of the proceedings of some of these events. He has been involved in a few international research projects with the Arab world: at the moment, he is in charge of a 2nd research project funded by the Spanish Association of International Cooperation AECI between Spain and Morocco on Arabic question answering.