Arabic Named Entity Recognition: Using Features Extracted from Noisy Data

Yassine Benajiba[1] **Imed Zitouni**[2] **Mona Diab**[1] **Paolo Rosso**[3]
[1] Center for Computational Learning Systems, Columbia University
[2] IBM T.J. Watson Research Center, Yorktown Heights
[3] Natural Language Engineering Lab. - ELiRF, Universidad Politécnica de Valencia
{ybenajiba,mdiab}@ccls.columbia.edu, izitouni@us.ibm.com, prosso@dsic.upv.es

Abstract

Building an accurate Named Entity Recognition (NER) system for languages with complex morphology is a challenging task. In this paper, we present research that explores the feature space using both gold and bootstrapped noisy features to build an improved highly accurate Arabic NER system. We bootstrap noisy features by projection from an Arabic-English parallel corpus that is automatically tagged with a baseline NER system. The feature space covers lexical, morphological, and syntactic features. The proposed approach yields an improvement of up to 1.64 F-measure (absolute).

1 Introduction

Named Entity Recognition (NER) has earned an important place in Natural Language Processing (NLP) as an enabling process for other tasks. When explicitly taken into account, research shows that it helps such applications achieve better performance levels (Babych and Hartley, 2003; Thompson and Dozier, 1997). NER is defined as the computational identification and classification of Named Entities (NEs) in running text. For instance, consider the following text:

Barack Obama is visiting the Middle East.

A NER system should be able to identify *Barack Obama* and *Middle East* as NEs and classify them as *Person* (PER) and *Geo-Political Entity* (GPE), respectively. The class-set used to tag NEs may vary according to user needs. In this research, we adopt the Automatic Content Extraction (ACE) 2007 nomenclature[1].

According to (Nadeau and Sekine, 2007), optimization of the feature set is the key component in enhancing the performance of a global NER system. In this paper we investigate the possibility of building a high performance Arabic NER system by using a large space of available feature sets that go beyond the explored shallow feature sets used to date in the literature for Arabic NER. Given current state-of-the-art syntactic processing of Arabic text and the relative small size of manually annotated Arabic NER data, we set out to explore a main concrete research goal: to fully exploit the level of advancement in Arabic lexical and syntactic processing to explore deeper linguistic features for the NER task. Realizing that the gold data available for NER is quite limited in size especially given the diverse genres in the set, we devise a method to bootstrap additional instances for the new features of interest from noisily NER tagged Arabic data.

2 Our Approach

We use our state-of-the-art NER system described in (Benajiba et al., 2008) as our baseline system (BASE) since it yields, to our knowledge, the best performance for Arabic NER . BASE employs Support Vector Machines (SVMs) and Conditional Random Fields (CRFs) as Machine Learning (ML) approaches. BASE uses lexical, syntactic and morphological features extracted using highly accurate automatic Arabic POS-taggers. BASE employs a multi-classifier approach where each classifier is tagging a NE class separately. The feature selection is performed by using an incremental approach selecting the top n features (the features are ranked according to their individual impact) at each iteration and keeping the set that yields the best results. In case of conflict - a word is classified with more than one class/tag simultaneously - the global NER system selects the output of the classifier with the highest precision.

The following is the feature set used in (Benajiba et al., 2008) and accordingly in the BASE system. **1. Context:** a $-/+1$ token window; **2. Lexical:** character $n-grams$ where n ranges from $1-3$; **3. Gazetteers:** automatically harvested and manually cleaned Person NE class (PER), Geopolitical Entity NE class (GPE), and Organization NE class (ORG) lexica; **4. POS-tag and Base Phrase Chunk (BPC)**: automatically tagged using AMIRA (Diab et al., 2007) which yields F-measures for both tasks in the high 90's; **5. Morphological features:** automatically tagged using the Morphological Analysis and Disambiguation for Arabic (MADA) tool to extract information about gender, number, person, definiteness and as-

[1] http://www.nist.gov/speech/tests/ace/index.htm

Proceedings of the ACL 2010 Conference Short Papers, pages 281–285,
Uppsala, Sweden, 11-16 July 2010. ©2010 Association for Computational Linguistics

pect for each word (Habash and Rambow, 2005);
6. Capitalization: derived as a side effect from running MADA. MADA chooses a specific morphological analysis given the context of a given word. As part of the morphological information available in the underlying lexicon that MADA exploits. As part of the information present, the underlying lexicon has an English gloss associated with each entry. More often than not, if the word is a NE in Arabic then the gloss will also be a NE in English and hence capitalized.

We devise an extended Arabic NER system (EXTENDED) that uses the same architecture as BASE but employs additional features to those in BASE. EXTENDED defines new additional syntagmatic features.

We specifically investigate the space of the surrounding context for the NEs. We explore generalizations over the kinds of words that occur with NEs and the syntactic relations NEs engage in. We use an off-the-shelf Arabic syntactic parser. State-of-the-art for Arabic syntactic parsing for the most common genre (with the most training data) of Arabic data, newswire, is in the low 80%s. Hence, we acknowledge that some of the derived syntactic features will be noisy.

Similar to all supervised ML problems, it is desirable to have sufficient training data for the relevant phenomena. The size of the manually annotated gold data typically used for training Arabic NER systems poses a significant challenge for robustly exploring deeper syntactic and lexical features. Accordingly, we bootstrap more NE tagged data via projection over Arabic-English parallel data. The role of this data is simply to give us more instances of the newly defined features (namely the syntagmatic features) in the EXTENDED system as well as more instances for the Gazetteers and Context features defined in BASE. It is worth noting that we do not use the bootstrapped NE tagged data directly as training data with the gold data.

2.1 Syntagmatic Features

For deriving our deeper linguistic features, we parse the Arabic sentences that contain an NE. For each of the NEs, we extract a number of features described as follows:

- Syntactic head-word (SHW): The idea here is to look for a broader **relevant** context. Whereas the feature lexical n-gram context feature used in BASE, and hence here for EXTENDED, considers the linearly adjacent neighboring words of a NE, SHW uses a parse tree to look at farther, yet related, words. For instance, in the Arabic phrase “SrH Ams An bArAk AwbAma ytrAs”, which means “declared yesterday that Barack Obama governs ...”, glossed “SrH/*declared* Ams/*yesterday* An/*that* bArAk/*Barack* AwbAmA/*Obama* ytrAs/*governs* ...”, is parsed in Figure 1. According to the phrase structure parse, the first parent sub-tree headword of the NE “bArAk AwbAmA” is the verb ‘ytrAs’ (governs), the second one is ‘An’ (that) and the third one is the verb ‘SrH’ (declared). This example illustrates that the word “Ams” is ignored for this feature set since it is not a syntactic head. This is a lexicalized feature.

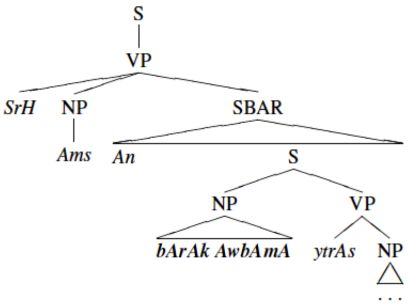

Figure 1: Example for the head word and syntactic environment feature

- Syntactic Environment (SE): This follows in the same *spirit* as SHW, but expands the idea in that it looks at the parent non-terminal instead of the parent head word, hence it is not a lexicalized feature. The goal being to use a more abstract representation level of the context in which a NE appears. For instance, for the same example presented in Figure 1, the first, second, and third non-terminal parents of the NE “bArAk AwbAmA” are ‘S’, ‘SBAR’ and ‘VP’, respectively.

In our experiments we use the Bikel implementation (Bikel, 2004) of the Collins parser (Collins, 1999) which is freely available on the web[2]. It is a head-driven CFG-style parser trained to parse English, Arabic, and Chinese.

2.2 Bootstrapping Noisy Arabic NER Data

Extracting the syntagmatic features from the training data yields relatively small number of instances. Hence the need for additional tagged data. The new Arabic NER tagged data is derived via projection exploiting parallel Arabic English data. The process depends on the availability of two key components: a large Arabic English parallel corpus that is sentence and word aligned, and a robust high performing English NER system. The process is as follows. We NE tag the

[2] http://www.cis.upenn.edu/∼dbikel/software.html#stat-parser

English side of the parallel corpus. We project the automatically tagged NER tags from the English side to the Arabic side of the parallel corpus. In our case, we have access to a large manually aligned parallel corpus, therefore the NER projection is direct. However, the English side of the parallel corpus is not NER tagged, hence we use an off-the-shelf competitive robust automatic English NER system which has a published performance of 92% (Zitouni and Florian, 2009). The result of these two processes is a large Arabic NER, albeit noisy, tagged data set. As mentioned earlier this data is used only for deriving additional instances for training for the syntagmatic features and for the context and gazetteer features.[3] Given this additional source of data, we changed the lexical features extracted from the BASE to the EXTENDED. We added two other lexical features: CBG and NGC, described as follows: - Class Based Gazetteers (CBG): This feature focuses on the surface form of the NEs. We group the NEs encountered on the Arabic side of the parallel corpus by class as they are found in different dictionaries. The difference between this feature and that in BASE is that the Gazetteers are not restricted to Wikipedia sources.

- N-gram context (NGC): Here we disregard the surface form of the NE, instead we focus on its lexical context. For each n, where n varies from 1 to 3, we compile a list of the $-n$, $+n$, and $-/+n$ words surrounding the NE. Similar to the CBG feature, these lists are also separated by NE class. It is worth highlighting that the NCG feature is different from the Context feature in BASE in that the window size is different $+/-1-3$ for EXTENDED versus $+/-1$ for BASE.

3 Experiments and Results

3.1 Gold Data for training and evaluation

We use the standard sets of ACE 2003, ACE 2004 and ACE 2005.[4] The ACE data is annotated for many tasks: Entity Detection and Tracking (EDT), Relation Detection and Recognition (RDR), Event Detection and Recognition (EDR). All the data sets comprise *Broadcast News* (BN) and *Newswire* (NW) genres. ACE 2004 includes an additional NW data set from the Arabic TreeBank (ATB). ACE 2005 includes a different genre of *Weblogs* (WL). The NE classes adopted in the annotation of the ACE 2003 data are: Person (PER), Geo Political Entity (GPE), Organization (ORG) and Facility (FAC). Additionally for the ACE 2004 and 2005 data, two NE classes are added to the ACE 2003 tag-set: Vehicles (e.g. Rotterdam Ship) and Weapons (e.g. Kalashnikof). We use the same split for train, development, and test used in (Benajiba et al., 2008).

3.2 Parallel Data

Most of the hand-aligned Arabic-English parallel data used in our experiments is from the Language Data Consortium (LDC).[5]. Another set of the parallel data is annotated in-house by professional annotators. The corpus has texts of five different genres, namely: newswire, news groups, broadcast news, broadcast conversation and weblogs corresponding to the data genres in the ACE gold data. The Arabic side of the parallel corpus contains 941,282 tokens. After projecting the NE tags from the English side to the Arabic side of the parallel corpus, we obtain a total of 57,290 Arabic NE instances. Table 1 shows the number of NEs for each class.

Class	Number of NEs	Class	Number of NEs
FAC	998	PER	17,964
LOC	27,651	VEH	85
ORG	10,572	WEA	20

Table 1: Number of NEs per class in the Arabic side of the parallel corpus

3.3 Individual Feature Impact

Across the board, all the features yield improved performance. The highest obtained result is observed where the first non-terminal parent is used as a feature, a Syntactic Environment (SE) feature, yielding an improvement of up to 4 points over the baseline. We experiment with different sizes for the SE, i.e. taking the first parent versus adding neighboring non-terminal parents. We note that even though we observe an overall increase in performance, considering both the {first, second} or the {first, second, and third} non-terminal parents decreases performance by 0.5 and 1.5 F-measure points, respectively, compared to considering the first parent information alone. The head word features, SHW, show a higher positive impact than the lexical context feature, NGC. Finally, the Gazetteer feature, CBG, impact is comparable to the obtained improvement of the lexical context feature.

3.4 Feature Combination Experiments

Table 2 illustrates the final results. It shows for each data set and each genre the F-measure obtained using the best feature set and ML approach. It shows results for both the dev and test data using the optimal number of features selected from

[3] Therefore, we did not do the full feature extraction for the other features described in BASE for this data.

[4] http://www.nist.gov/speech/tests/ace/

[5] All the LDC data are publicly available

		ACE 2003		*ACE 2004*			*ACE 2005*		
		BN	*NW*	*BN*	*NW*	*ATB*	*BN*	*NW*	*WL*
	FreqBaseline	*73.74*	*67.61*	*62.17*	*51.67*	*62.94*	*70.18*	*57.17*	*27.66*
dev	*All-Synt.*	83.41	79.11	76.90	**72.90**	74.82	81.42	**76.07**	54.49
	All	**83.93**	**79.72**	**78.54**	72.80	**74.97**	**81.82**	75.92	**55.65**
test	*All-Synt.*	83.50	78.90	76.70	**72.40**	73.50	81.31	75.30	57.30
	All	**84.32**	**79.4**	**78.12**	72.13	**74.54**	**81.73**	**75.67**	**58.11**

Table 2: Final Results obtained with selected features contrasted against all features combined

the all the features except the syntagmatic ones (`All-Synt.`) contrasted against the system including the semantic features, i.e. All the features, per class *All* . The baseline results, *FreqBaseline*, assigns a test token the most frequent tag observed for it in the gold training data, if a test token is not observed in the training data, it is assigned the most frequent tag which is the O tag.

4 Results Discussion

Individual feature impact results show that the syntagmatic features are helpful for most of the data sets. The highest improvements are obtained for the 2003 BN and 2005 WL data-sets. The improvement varies significantly from one data-set to another because it highly depends on the number of NEs which the model has not been able to capture using the contextual, lexical, syntactic and morphological features.

Impact of the features extracted from the parallel corpus per class: The syntagmatic features have varied in their influence on the different NE classes. Generally, the LOC and PER classes benefitted more from the head word features, SHW), than the other classes. On the other hand for the syntactic environment feature (SE), the PER class seemed not to benefit much from the presence of this feature. *Weblogs*: Our results show that the random contexts in which the NEs tend to appear in the WL documents stand against obtaining a significant improvement. Consequently, the features which use a more global context (syntactic environment, SE, and head word, SHW, features) have helped obtain better results than the ones which we have obtained using local context namely CBG and NGC.

5 Related Work

Projecting explicit linguistic tags from another language via parallel corpora has been widely used in the NLP tasks and has proved to contribute significantly to achieving better performance. Different research works report positive results when using this technique to enhance WSD (Diab and Resnik, 2002; Ng et al., 2003). In the latter two works, they augment training data from parallel data for training supervised systems. In (Diab, 2004), the author uses projections from English into Arabic to bootstrap a sense tagging system for Arabic as well as a seed Arabic WordNet through projection. In (Hwa et al., 2002), the authors report promising results of inducing Chinese dependency trees from English. The obtained model outperformed the baseline. More recently, in (Chen and Ji, 2009), the authors report their comparative study between monolingual and cross-lingual bootstrapping. Finally, in Mention Detection (MD), a task which includes NER and adds the identification and classification of nominal and pronominal mentions, (Zitouni and Florian, 2008) show the impact of using a MT system to enhance the performance of an Arabic MD model. The authors report an improvement of up to 1.6F when the baseline system uses lexical features only. Unlike the work we present here, their approach requires the availability of an accurate MT system which is a more expensive process.

6 Conclusion and Future Directions

In this paper we investigate the possibility of building a high performance Arabic NER system by using lexical, syntactic and morphological features and augmenting the model with deeper lexical features and more syntagmatic features. These extra features are extracted from noisy data obtained via projection from an Arabic-English parallel corpus. Our results show that we achieve a significantly high performance for almost all the data-sets. The greatest impact of the syntagmatic features (1.64 points of F-measure) is obtained for the ACE 2004, BN genre. Also, the WL genre yields an improvement of 1.16 F1 points absolute.

Acknowledgments

This work has been partially funded by DARPA GALE project. The research of the last author was funded by MICINN research project TEXT-ENTERPRISE 2.0 TIN2009-13391-C04-03 (Plan I+D+i).

References

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

Y. Benajiba, M. Diab, and P. Rosso. 2008. Arabic named entity recognition using optimized feature sets. In *Proceedings of EMNLP'08*, pages 284–293.

Daniel M. Bikel. 2004. *On the parameter space of generative lexicalized statistical parsing models*. University of Pennsylvania, Philadelphia, PA, USA. Supervisor-Marcus, Mitchell P.

Z. Chen and H. Ji. 2009. Can one language bootstrap the other: A case study of event extraction. In *Proceedings of NAACL'09*.

M. Collins. 1999. *Head-Driven Statistical Models for Nat- ural Language Parsing*. University of Pennsylvania, Philadelphia, PA, USA.

Mona Diab and Philip Resnik. 2002. An unsupervised method for word sense tagging using parallel corpora. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, pages 255–262, Philadelphia, Pennsylvania, USA, July. Association for Computational Linguistics.

M. Diab, K. Hacioglu, and D. Jurafsky, 2007. *Arabic Computational Morphology: Knowledge-based and Empirical Methods*, chapter 9. Springer.

Mona Diab. 2004. Bootstrapping a wordnet taxonomy for arabic. In *Proceedings of First Arabic Language Technology Conference (NEMLAR), Cairo Egypt,*.

N. Habash and O. Rambow. 2005. Arabic Tokenization, Part-of-Speech Tagging and Morphological Disambiguation in One Fell Swoop. In *Proc. of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 573–580, Ann Arbor, Michigan, June. Association for Computational Linguistics.

R. Hwa, P. Resnik, and A. Weinberg. 2002. Breaking the resource bottleneck for multilingual parsing. In *In Proceedings of the Workshop on Linguistic Knowledge Acquisition and Representation: Bootstrapping Annotated Language Data*.

D. Nadeau and S. Sekine. 2007. A Survey of Named Entity Recognition and Classification. *Linguisticae Investigationes*, 30(7).

H.-T. Ng, B. Wang, and Y.-S. Chan. 2003. Exploiting parallel texts for word sense disambiguation: An empirical study. In *ACL'03*, pages 455–462, Sapporo, Japan.

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

I. Zitouni and R. Florian. 2008. Mention detection crossing the language barrier. In *Proceedings of EMNLP'08*, Honolulu, Hawaii, October.

Imed Zitouni and Radu Florian. 2009. Cross language information propagation for arabic mention detection. *Journal of ACM Transactions on Asian Language Information Processing*, December.