Named Entity Recognition in Arabic: A Combined Approach

June 4, 2009
David O'Steen, David Breeden
Final Project
CS 224N / Ling 237

Abstract

The problem of Named Entity Recognition (NER) in Arabic is both difficult and widely applicable. The challenges inherent in agglutinative languages like Arabic demand specialized methods beyond currently-understood language-independent learning algorithms. This paper presents an SVM-based approach for Arabic NER with language-generic and language-specific features, resulting in a 10-30 point increase in F1 score over baseline for person, location and organization named entity categories.

1 Introduction

Named Entity Recognition (NER) is an important gateway to higher levels of textual semantic understanding, since most important semantic relationships constitute associations between named entities. Thus, extracting named entities is integral to applications such as question answering, and also beneficial feedback to lower-level applications like machine translation or optical character recognition.

For these reasons, NER has flourished under a great deal of research for the last 15 years. However, most of this research has been focused in Latin languages, with tools and corpora for Eastern languages lagging behind. In addition, the inherently distinct nature of such languages makes it difficult to directly port methods for well-studied languages to immediate use in novel domains.

The Arabic language in particular presents significant obstacles and opportunities, due to its wide global usage and its intricate morphology. This paper implements a NER system for Arabic that incorporates language-generic methods borrowed from established language-independent NLP systems with language-specific morphological analysis built up from Arabic grammar rules.

The paper is organized as follows: Section 2 presents recent work in Arabic

NER. Section 3 enumerates some of the facets of the Arabic language that require a model tailored to the structure and dynamics of Arabic grammar. Section 4 describes the components of an SVM-based approach to accomplish Arabic NER. Section 5 presents an evaluation of this method and its constituent features. Section 6 presents a discussion of the subsequent results, and Section 7 provides some concluding remarks.

2 Related Work

The area of Arabic NER has only recently emerged, thanks to the ongoing development of Arabic language processing tools and corpora. While this data has been slow to materialize, many investigators have leveraged the large amount of research in language-independent NER to achieve moderate success in the short period of time where sufficient corpora have been available. This success has mainly been achieved by a combination of language-independent discriminative models and morphological analysis built on grammatical rules specific to Arabic.

Early work in Arabic NER focused on the intricate morphology of the Arabic language. Unlike many commonly studied languages in natural language processing (NLP), Arabic is highly inflectional and early methods were justified to rely on rule-based schemes and Arabic-specific morphological analysis [7]. The ambiguities encountered in Arabic text would be extremely difficult to navigate without such domain-specific knowledge, as illustrated in Section 3. More recently, Shaalan [8] demonstrated the effectiveness of a similarly rule-based system, backed by heroic amounts of data, for automating data collection.

However, these rule-based methods have benefited from the more robust models developed in the field of language-independent NER. Namely, Benajiba *et al.* have investigated the use of maximum entropy models [3], conditional random fields [4] and SVMs [1] for application to Arabic NER, integrating language-specific systems such as the Morphological Analysis and Disambiguation for Arabic (MADA) tool [6]. With a combination of these methods, they achieve an overall F1 score of 83.5 on the ACE 2003 data, which represents the state of the art in Arabic NER [2].

3 The Arabic Language

Many features of the Arabic language make text processing difficult. The first of these that is immediately noticeable to any NLP investigator used to Latin languages is the absence of uppercase letters. Arabic possesses no useful shape features, which are often very informative in other languages.

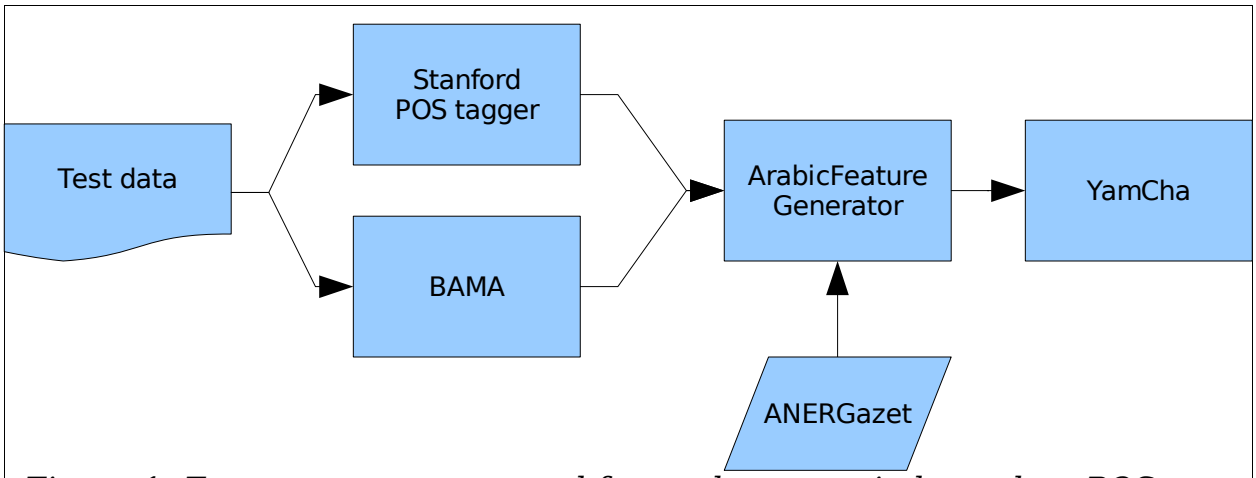

Figure 1: Features are aggregated from a language-independent POS tagger, morphological analyzer, and gazetteers

The second main difficulty is Arabic's heavy use of inflection. Because of this, Arabic words often take prefixes and suffixes, such that one contiguous token can contain 3 (or more) clitics. This results in very sparse data which can lead to overtraining and poor generalization without massive amounts of data. There are two standard methods to solve this problem. The first is to discard the affixes and reduce the data to its stems. This entails sacrificing a great deal of the contextual information given by a word's affixes. The second (preferred) method is to segment each word and thenceforth treat its clitics as distinct tokens. Both of these methods require segmenting each word into its clitics, which is a far from trivial problem. For these reasons, feasible approaches to NER in the Arabic domain necessitate language-specific analysis.

4 Methods

Our approach combined a number of publicly available systems and corpora (Figure 1). At the highest level, however, we used the YamCha[1] tool to reduce the NER task to text chunking on IOB data. The primary challenge then becomes defining suitable features to discriminate named entity sequences. To this end, our system makes use of the Buckwalter Arabic Morphological Analyzer (BAMA)[2] and the Stanford POS tagger [9, 10] as well as the ANERgazet[3] battery of gazetteers [3]. BAMA is widely used as the first layer for morphological analysis in a multitude of Arabic systems across NLP areas, although we also considered the Linguistica system [5] for unsupervised morphological analysis. The Stanford POS tagger was selected for its demonstrated good performance on language-independent POS tagging, and the convenience of its pre-trained Arabic tagger. Though

[1] http://chasen.org/~taku/software/yamcha/
[2] http://www.qamus.org/morphology.htm
[3] http://users.dsic.upv.es/~ybenajiba/

the modules differ, this design closely follows the work of Benajiba 2008 [1].

4.1 Language-independent features

To take advantage of language-independent dependencies on named entities, we included the following word features, which are agnostic to the target language:

Context
Our baseline feature set are the context features defaulted to by YamCha: the two previous and subsequent tokens, and the two previous inferred categories. This encapsulates quite a lot of information, particularly in the determination of extending or terminating named entities.

Token length
Although it is a very general feature, the length of a token would appear to capture some useful information, since named entities, particularly location and organization names, tend to be longer than non-entities. The descriptiveness of this feature is explored at greater depth in Section 6.

Presence in gazetteers
ANERgazet consists of three distinct name lists for people, locations and organizations, respectively. Most of these names were culled from Wikipedia. The presence of a word in one of these gazetteers indicates a strong likelihood that the word should be given the corresponding category.

Part-of-Speech
We used the Arabic Stanford POS tagger[4] to provide the likely part-of-speech (POS) for each word. This tagger does not incorporate any language-specific methods, but was trained on Arabic Treebank (ATB) train data and demonstrated 96.72% accuracy on the dev set (77.49% on unknown words).

4.2 Arabic-specific features

Because of the highly inflectional properties of Arabic described above, morphological analysis is a necessary component of any Arabic NLP system. As such, we derive a set of Arabic-specific features for each word in our test data using the BAMA. BAMA is a simple Arabic morphological analyzer which uses a rule-based system to propose possible segmentations of Arab words into stems, prefixes and suffixes. The model leverages extensive dictionaries relating clitics to rich grammatical information. Our system uses BAMA to segment each word in the data into a prefix, stem and suffix, and then generate POS, gender, person, number and case features for each.

[4] http://nlp.stanford.edu/software/tagger.shtml

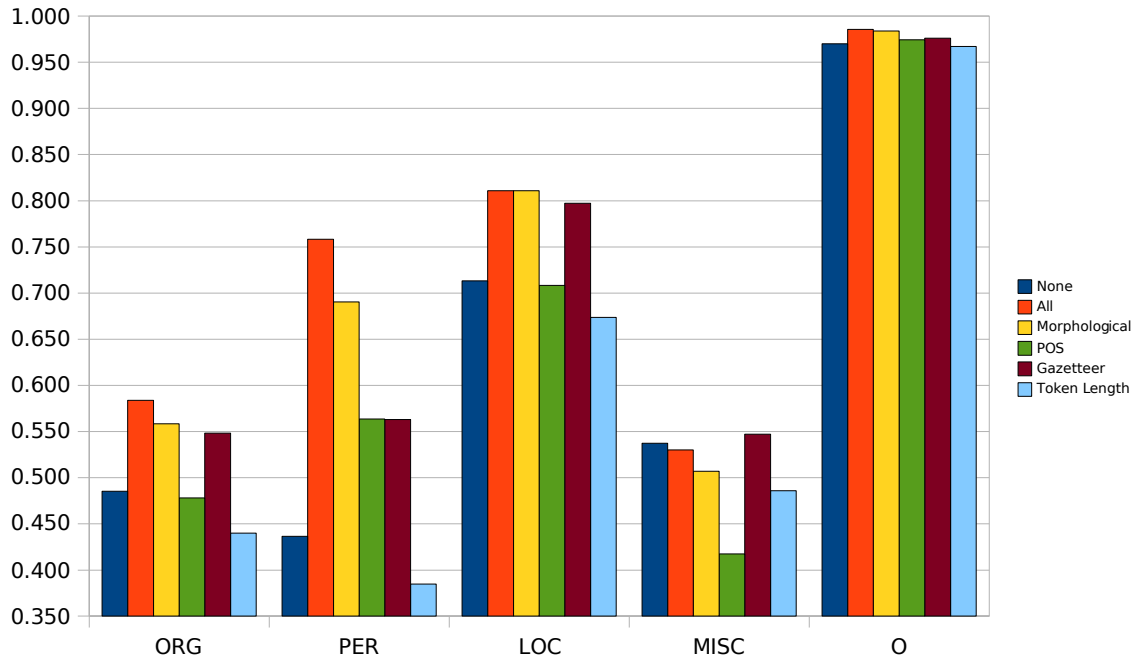

Figure 2: F1 scores for Various Features

BAMA makes the assumption that all Arabic words can be segmented thusly (although affixes may be null), and that they satisfy particular length constraints (prefixes are constrained to be no longer than 4 characters, suffixes no longer than 6).

The specific form of output BAMA provides is a set of valid segmentations for each given word, along with a POS tag encoded with these grammatical features. While this tag contains all of the information listed above, it is advantageous to split this information into separate features to reduce data sparsity. By modifying BAMA, we reduced the set of distinct POS tags used in the BAMA dictionaries from 634 to 104 (an 83.6% reduction). Considering that there are only a few dozen basic tags defined for the Arabic Treebank, it is apparent that there are more features that could be derived from the parsed tags (for instance, our system does not extract definiteness or mood, which are encoded in the BAMA POS tags).

Contrary to other approaches, we neither neglect the affixes nor explicitly segment each word in preprocessing, but rather annotate each word with a set of its proposed affixes and their corresponding grammatical features. This approach maintains the large amount of contextual information contained in and surrounding each word, capturing more inter-word dependency.

5 Experiments & Results

	Precision	Recall	F1
ORG	81.4%	45.5%	58.4%
PER	82.6%	70.0%	75.8%
LOC	84.7%	77.7%	81.1%
MISC	86.3%	38.2%	53.0%
O	97.5%	99.6%	98.5%

Table 1: Performance with All Features

		Actual				
		ORG	PER	LOC	MISC	O
Predicted	ORG	81.4%	4.8%	2.8%	5.2%	5.8%
	PER	3.6%	82.6%	3.6%	2.6%	7.5%
	LOC	5.7%	3.4%	84.7%	1.5%	4.8%
	MISC	3.4%	1.2%	2.1%	86.3%	7.0%
	O	0.7%	0.6%	0.6%	0.5%	97.5%

Table 2: Confusion Matrix for All Features Enabled Enabled

We evaluated our system by measuring precision, recall and F1 scores with a variety of distinct feature sets. For each experiment, we split the ANERcorp dataset (consisting of 150,285 tokens) [3] into training and test sets of equal size. We used as our baseline the contextual features described above. We then tested performance with each of our additional features in turn, and finally with all features. Figure 2 shows the results for each of the feature sets.

Generally, we found that including additional features improved the performance of the model, though, their relative effectiveness varied among the named entity classes. The improvements in the F1 scores from the baseline model are a result of increasing recall at the cost of precision. The baseline model has a very high precision (~90%), whereas its recall is generally poor (ranging from ~30% to ~60% for entities excluding O). This indicates that the baseline is rather conservative in assigning labels, and that addition of other features improves the model by increasing its coverage.

Our most significant performance gains were for person entities. Unlike the other entities, it appears that the performance gain was only achievable through a combination of multiple features, since no one feature was able to come close the performance of the model with all features enabled. Looking deeper into the data, it appears that the gazetteer model has high precision and low recall (90% and 41%, respectively), whereas the morphological model has lower precision, but higher recall (83% and 59%, respectively). Since the two features have relatively distinct strengths for

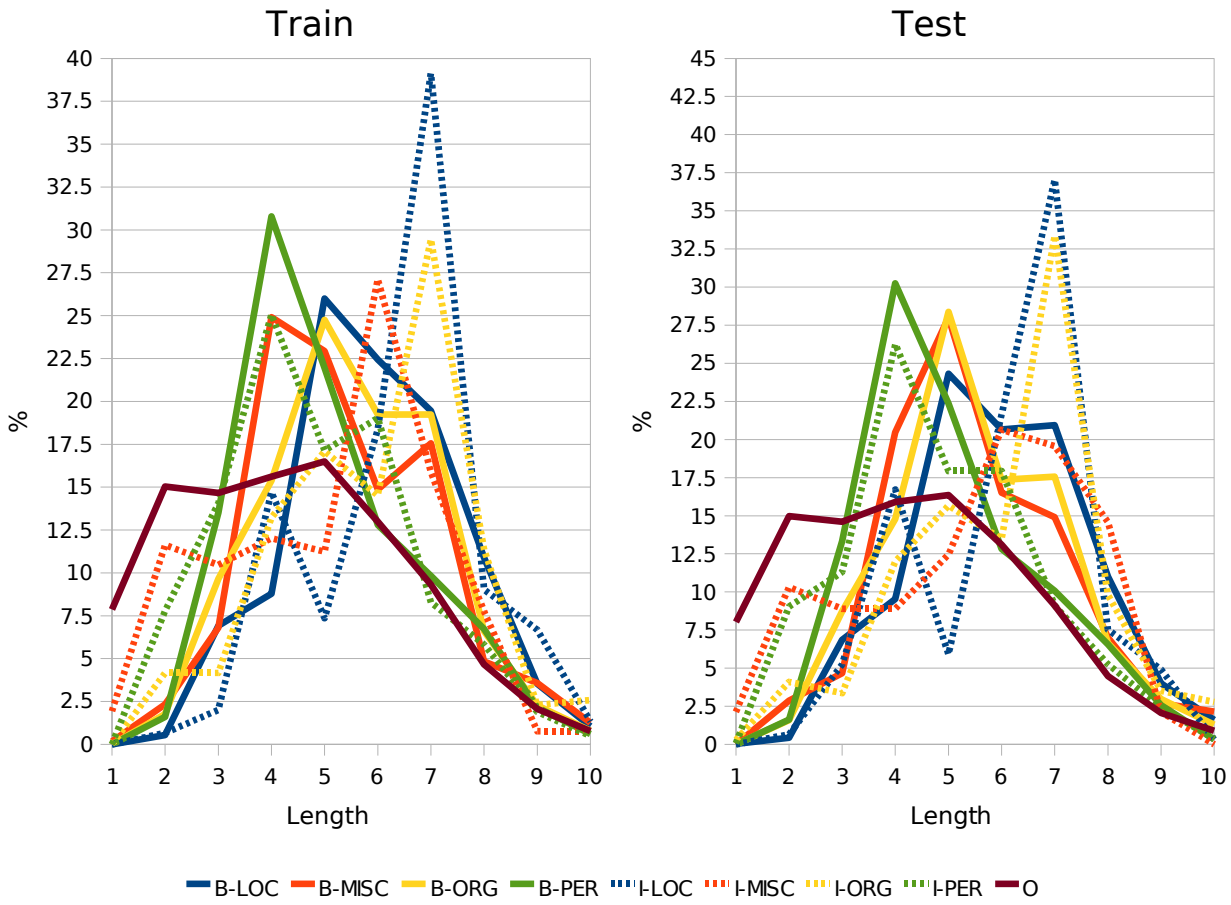

Figure 3: Word length distributions for train and test splits

this entity class, they are more effective when combined.

Unfortunately, none of our features seemed to improve performance for categorizing the miscellaneous entities. In fact, the baseline contextual model outperformed the morphological features model for this entity class. Given the expected heterogeneity of a catch-all miscellaneous category, it makes sense that morphological features may be less useful. When we examined the data more closely, we found a very high recurrence of two Arabic words that roughly correspond to “dollar”. It follows that a more simple contextual model may capture instances like this better, as the morphological model may find “false leads” when extracting features. Though we did not have a gazetteer of miscellaneous entities, it is likely that such a corpus would aid recognition of a diverse group of entities.

Tables 1 and 2 show the scores and confusion matrix, respectively, for the model with full set of features. While these results are significantly better than the baseline, there is still a large gap between the precision and recall. This implies that further gains in performance could be achieved by increasing the coverage, perhaps through the use of larger gazetteers, or

richer morphological features.

6 Discussion

To hypothesize the discriminative content of the word length feature, Figure 3 plots the distributions of word lengths for each named entity category in sample train and test splits. A number of patterns are apparent from this chart. The distributions of word lengths for named entities vs. other words are ostensibly different. Non-entities exhibit a smoother distribution with generally fewer letters per word, while named entities generally have more letters per word. Further, person names have fewer letters than organization, location or other names. Finally, the inside words of locations, organizations and miscellaneous entities appear to be generally longer than their beginning words. However, as observed in our experiments, the word length feature actually hurt performance in all areas. This would suggest that including this feature could lead to overtraining. It is possible that the slight differences in the two charts account for some generalization error.

In addition to exploring the word length feature, we also decided to examine the gazetteers in greater detail. We were initially surprised at the effectiveness of these features, given their simplicity. To verify our results were reasonable, we decide to test a primitive lookup model using only the gazetteers. To classify each word, we looked it up in each of the gazetteers, and if it was in one, we labeled it accordingly. The scores obtained from this process are shown in Table 3. Like the other models we evaluated, this one has a high precision relative to its recall.

This makes sense for the gazetteers; given that entities tend to have distinct names, the process of looking up them up is likely to be accurate. Yet having a comprehensive list to give good coverage is difficult because of the large volume of possible names. It also makes sense that organizations would be the weakest of these three categories. Though our particular gazetteer is small for this entity group, it is still a class that is more likely to be composed of multiple tokens, many of which will be regular words.

	Precision	Recall	F1
PER	87.1%	38.4%	53.3%
LOC	85.9%	43.5%	57.7%
ORG	56.3%	17.6%	26.8%

Table 3: Scores for Basic Gazetteer Lookup Model

7 Conclusions

This system could improve along a number of dimensions. First, BAMA could be overlaid with a system for discriminating among possible clitic segmentations. In the current system,
disambiguation is done arbitrarily, although for many words there are several possible segmentations. A system like MADA [6] would provide a sound method for this, and provide better morphological features. For this task, MADA might also render the Stanford POS tagger irrelevant, since it produces POS tags that incorporate Arabic morphology, instead of learning from raw ATB data.

Despite these areas for improvement, the system and results presented in this paper achieve satisfactory results, showing marked improvement with the features presented. Importantly, they display the value of domain-specific information in addition to general methods for Arabic NER and provide a map of the special considerations that ought to be taken into account when entering a new domain with idiosyncratic demands.

8 References

1.Benajiba, Y., Diab, M., Rosso, P.: Arabic Named Entity Recognition: An SVM-based approach. In *Proceedings of the International Arab Conference on Information Technology*, 2008.
2.Benajiba, Y., Diab, M., Rosso, P.: Arabic Named Entity Recognition using Optimized Feature Sets. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, 2008.
3.Benajiba, Y., Rosso, P.: ANERsys 2.0: Conquering the NER Task for the Arabic Language by Combining the Maximum Entropy with POS-tag Information. In *Proceedings of the Workshop on Natural Language-Independent Engineering, 3rd Indian International Conference on Artificial Intelligence*, 2007.
4.Benajiba, Y., Rosso, P.: Arabic Named Entity Recognition using Conditional Random Fields. In *Proceedings of the Workshop on HLT & NLP within the Arabic World*, 2008.
5.Goldsmith, J.: Linguistica: An Automatic Morphological Analyzer. In *Proceedings from the Main Session of the Chicago Linguistic Society's Thirty-sixth Meeting*, 2000.
6.Habash, N., Rambow, O.: Arabic Tokenization, Part-Of-Speech Tagging and Morphological Disambiguation in One Fell Swoop. In *Proceedings of the Workshop of Computational Approaches to Semitic Languages*, 2005.
7.Maloney, J., Niv, M.: TAGARAB: A Fast, Accurate Arabic Name Recognizer Using High-Precision Morphological Analysis. In *Proceedings of the Workshop on Computational Approaches to Semetic Languages*, 1998.
8.Shaalan, K. F., Raza, H.: Arabic Named Entity Recognition from Diverse

Text Types. In *Proceedings of the 6th International Conference on Advances in Natural Language Processing*, 2008.
9.Toutanova, K., Klein, D., Manning, C., Singer, Y.: Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network. In *Proceedings of HLT-NAACL 2003*, 2003.
10.Toutanova, K., Manning, C. D.: Enriching the Knowledge Sources Used In a Maximum Entropy Part-of-Speech Tagger. In *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, 2000.