Improving NER in Arabic Using a Morphological Tagger

Benjamin Farber†, Dayne Freitag†, Nizar Habash‡, Owen Rambow‡

†Fair Isaac Corporation
{BenFarber,DayneFreitag}@fairisaac.com
‡Center for Computational Learning Systems, Columbia University
{habash,rambow}@ccls.columbia.edu

Abstract

We discuss a named entity recognition system for Arabic, and show how we incorporated the information provided by MADA, a full morphological tagger which uses a morphological analyzer. Surprisingly, the relevant features used are the capitalization of the English gloss chosen by the tagger, and the fact that an analysis is returned (that a word is not OOV to the morphological analyzer). The use of the tagger also improves over a third system which just uses a morphological analyzer, yielding a 14% reduction in error over the baseline. We conduct a thorough error analysis to identify sources of success and failure among the variations, and show that by combining the systems in simple ways we can significantly influence the precision-recall trade-off.

1. Introduction

For a range of applications of natural language processing—from machine translation, to question answering, to information retrieval—it is often beneficial to devote one stage of processing to the identification of names. Automated *named entity recognition* (NER) has been demonstrated on a number of languages, and has reached near-human levels of performance on English. Not so on Arabic. Not only has less annotated training data been created for this task in Arabic, but Arabic exhibits greater lexical ambiguity and morphological variety than English, and written Arabic lacks the overt clue of capitalization.

In this study we investigate whether a morphological tagger can mitigate these sources of difficulty in Arabic. In general, a written Arabic word may have dozens of mutually incompatible morphological analyses. This ambiguity is due in part to the orthography of conventional written Arabic, which does not require the inclusion of short vowels. Confronted with a word form in context, the task of a morphological tagger is to determine all morphological properties of the word. The process is often called simply "part-of-speech tagging" in English, because the set of tags to completely describe the set of morphological forms of the language is much smaller than in Arabic (about 50 tags in English against thousands in Arabic).

Many proposed approaches to English NER do not make use of part of speech. We claim that in Arabic the information returned by a morphological tagger is critical to maximize performance. Consistent with the state of the art, our Arabic NER system involves the application of models trained with name-annotated Arabic data. In this paper we explore one way in which the analysis of the morphological tagger can be integrated with such a system: The output of the tagger is converted into additional *features* of the input, which are made available to the learner responsible for producing the NER model. Our experiments show that the morphological tagger provides critical information in the form of lexical disambiguation.

The rest of the paper is organized as follows. Section 2. describes the training and configuration of our baseline NER system, which is essentially language-neutral. Section 3. explores in greater depth the problem of Arabic morphological analysis, and describes the morphological tools we use in these experiments. Section 4. describes how these tools are integrated in the NER system and presents the empirical evidence that this integration is beneficial. We end with an error analysis in Section 5., which prompts a discussion of future work in the final section, Section 6..

2. Named Entity Recognition

The current state of the art sees the problem of named entity recognition as one of sequence labeling (or "structured classification"), akin to problems like part-of-speech tagging. Given a stream of input words, typically a single sentence, the problem is to assign a label to each word in a way that unambiguously identifies any named entity mentions. If the object is to distinguish the names of one type of entity (e.g., person), the use of *BIO labeling* is common, and is the labeling scheme we employ here. In BIO labeling, the name-initial word is assigned a `B` label, any other name constituent an `I`, and all other words an `O`. This labeling generalizes in an obvious way to one that accounts for multiple entity types simultaneously.

All of the approaches that have been tried for problems like part-of-speech have also been applied to NER, including decision trees (Baluja et al., 1999), log-linear models (maximum entropy) (Borthwick, 1999), and support vector machines (Isozaki and Kazawa, 2002). However, inasmuch as these paradigms do not naturally address the sequential nature of the problem, they have largely given way to approaches that model labeling as a series of related decisions, in particular hidden Markov models (HMM) (Bikel et al., 1997), structured perceptrons (Collins, 2002), and conditional random fields (CRF) (McCallum and Li, 2003). The last two approaches are particularly appealing, because in contrast to generative approaches like HMMs, they make no unwarranted independence assumptions, and can therefore fruitfully incorporate arbitrary features of the input.

2.1. Structured Perceptron for NER

Let S be our input, a sequence of words, and S_i be a single word in the sequence. Our task is to produce a sequence L,

the same length as S, with each L_i from our set of candidate labels (in the simplest case $\{\texttt{b}, \texttt{i}, \texttt{o}\}$).

Our baseline NER system employs the structured perceptron proposed by Collins (2002), which exhibits the above mentioned flexibility, conceptual simplicity, and competitive performance. A model in this paradigm can be regarded as a collection of competing classifiers, one for each label type. At the core of the labeling problem is the decision what label to return for a given S_i. To the structured perceptron this decision is a function of *features* of the input, which are typically Boolean and fall into one of two categories. In one set, we have features that are sensitive to previous labeling decisions. There is typically one feature for every preceding distinct label context observed during training (e.g., L_{i-2} and L_{i-1} were `b` and `i`, respectively), within a user-specified context window.

The remaining features are arbitrary Boolean functions of S and i. For example, it is common to define features reflecting the identity of S_i (e.g., 1 if the current word is `the`, else 0), as well as of S_{i-1}, S_{i+1}, etc. If gazetteers or other lexical resources are available, features may be defined for the current or neighboring words reflecting their respective membership in the various lists. Similarly, features may reflect easily measurable characteristics of the current or neighboring words, such as capitalization.

2.2. Arabic NER

Before this model can be applied to Arabic, either for training or labeling, an input sentence must be segmented into a stream of words. We employ a naive tokenization, in which a "word" is an unbroken string of letters from the Arabic alphabet, an unbroken string of numbers, or a single punctuation character. Note that this tokenization policy is arguably sub-optimal for Arabic, which has a rich morphology and an orthography that agglutinates some clitics (Section 3.1.), which in English would treated as separate words. This fact, together with the rich inflectional morphology (Section 3.1.), contributes to a problem of data sparseness. We do not respond to this issue in this study, although our morphological tagger accounts for such phenomena, but see Section 6. for possible future directions in this area.

Our experimental data for this study is the newswire portion of the Arabic training data from Year 2005 of the Adaptive Content Extraction (ACE) program. Annotations in this data identify not only names, but also nominal and pronominal references to named entities; we restrict our attention only to the name mentions of all entity types distinguished in the data, the most frequent of which are persons, organizations, and geo-political entities (GPEs).

In order to train a structured perceptron on this data, all that remains is to define the features that will be visible to the model. We define word-identity features, as described above, for every position up to three tokens distant from the reference position on either side. Thus, observing the word في *fy*[1] 'in' at the current position triggers a different feature than observing it at the preceding position.

In addition to these word-identity features, we define features that reflect membership in word classes derived through a statistical analysis of a large repository of Arabic newswire articles (Arabic Gigaword). In this procedure, words are associated with probability distributions over other words observed to have occurred in their close context, then clustered in a way that heuristically minimizes information loss. Several studies have documented the utility of such features for NER (Freitag, 2004; Miller et al., 2004), and our informal experiments confirm their considerable benefit in the processing of Arabic. As with word identity, there are different cluster-membership features for every position relative to the reference position, up to a distance of three tokens. An example feature of this type might be "the word succeeding the current word is in Cluster 42."

3. Arabic Morphological Analysis and Disambiguation

3.1. Arabic Morphology

Arabic is a morphologically complex language with a large set of morphological features[2]. These features are realized using both concatenative (affixes and stems) and templatic (root and patterns) morphology with a variety of morphological and phonological adjustments that appear in word orthography and interact with orthographic variations. In addition to inflectional features such as gender, number, person and voice, Arabic has a set of clitics that are written attached to the word and thus increase its ambiguity. Clitics include (a.) conjunction proclitics such as و+ *w+* 'and' and ف+ *f+* 'then', (b.) particle proclitics such as ل+ *l+* 'to/for' and ب+ *b+* 'by/with', (c.) the definite article ال+ *Al+* 'the', and (d.) pronominal enclitics such as هم+ *+hm* 'their/them'. To model Arabic morphology, we use the BAMA morphological analyzer and the MADA system for morphological disambiguation.

3.2. BAMA

We use the Buckwalter Arabic Morphological Analyzer (BAMA) (Buckwalter, 2004), to obtain all possible word analyses. BAMA models Arabic morphology for over 35K lexemes. The number of possible fully diacritized forms per lexeme varies from 3K forms for nouns to 17K forms for verbs. Figure 1 shows three BAMA analyses of the word form بين *byn*. In all, BAMA judges that this word form has 19 possible analyses. For each analysis, the table provides the surface diacritization, the lexeme, the list of morphological features, the morpheme segmentations and the English gloss.

[1] All Arabic transliterations are provided in the Habash-Soudi-Buckwalter transliteration scheme (Habash et al., 2007). This scheme extends Buckwalter's transliteration scheme (Buckwalter, 2004) to increase its readability while maintaining the 1-to-1 correspondence with Arabic orthography as represented in standard encodings of Arabic, i.e., Unicode, CP-1256, etc. The following are the only differences from Buckwalter's scheme (which is indicated in parentheses): Ā ٱ(|), Â أ(>), ŵ ؤ (&), Ǎ إ(<), ŷ ىء (}), ħ ة (p), θ ث (v), ð ذ (*), š ش ($), Ď ظ (Z), ς ع (E), γ غ (g), ý ى (Y), ã ً (F), ũ ٌ (N), ĩ ٍ (K).

[2] Arabic words can be analyzed using up to fourteen features: POS, person, number, gender, voice, aspect, determiner proclitic, conjunctive proclitic, particle proclitic, and pronominal enclitic, nominal case, nunation, idafa (possessed), and mood.

```
;;WORD byn
bay~ana=[bay~an_1 POS:V +PV +S:3MS BW:+bay~an/PV+a/PVSUFF_SUBJ:3MS]=declare/demonstrate
bayonu=[bayona_1 POS:N +NOM +DEF BW:+bayon/NOUN+u/CASE_DEF_NOM]=between/among
biyn=[biyn_1 POS:PN BW:+biyn/NOUN_PROP+]=Ben
```

Figure 1: Three BAMA analyses for the word بين *byn*.

3.3. MADA

MADA, The Morphological Analysis and Disambiguation for Arabic tool, is an off-the-shelf resource for Arabic disambiguation (Habash and Rambow, 2005; Habash, 2007). MADA selects among BAMA analyses using a combination of classifiers that tag words on all 14 orthogonal dimensions of Arabic morphology. The version of MADA used in this paper was trained on the Penn Arabic Treebank (PATB) part 3 (Maamouri et al., 2004). Habash and Rambow (2005) report disambiguation accuracy over 96% (ignoring case, idafa, mood and nunation) and word-level PATB tokenization accuracy over 99.3%.

4. Incorporating Morphology into NER

An important product of MADA tagging, one which provides the most immediate benefit to NER, is full lexical disambiguation. In Figure 1, for example, one of the analyses indicates that the word بين *byn* may correspond to the English given name *Ben*. Thus, an accurate assessment by MADA that '*Ben*' is the correct analysis, out of the 19 possible, should serve as a powerful clue to the NER model that a name is present.

4.1. Enhanced NER Model

In order to pursue this intuition, we enhanced the baseline NER model by introducing two new feature classes dependent on the output of morphological analysis. One, *GlossCap*, is true in those cases where a word's gloss is capitalized. The other, *OOV*, is true if no entry exists for a word in our morphological database. Both of these feature types carry potentially important information for NER. GlossCap is of course a kind of substitute for the capitalization missing from written Arabic, which in English serves as an important clue. OOV signals the presence of an unfamiliar word; such words are often names, in Arabic as in English. We also experimented with other features returned by MADA, including the core part-of-speech (noun, verb, preposition, ...), but only these two features provided an improvement.

As with the other feature classes, we define separate features for every offset relative to a target position up to three tokens on either side. Thus, the enhanced model includes a total of fourteen new features, seven from each of the two new feature classes.

Note that one may think that the disambiguation provided by MADA may not be strictly necessary: perhaps the set of all possible analyses returned by BAMA alone may be sufficient. This hypothesis is motivated by the fact that the NER model performs a kind of disambiguation itself, and it thus might derive just as much benefit from the information that a word form has a *potential* interpretation as a name as that it has a *definitive* one.

	Base	BAMA	MADA
F1	0.667	0.676	0.715
Precision	0.714	0.713	0.735
Recall	0.627	0.643	0.697

Table 1: Summary F1, precision, and recall for three NER system variants.

We are thus interested in measuring the extent of the benefit provided by MADA's disambiguation of the BAMA analysis. We therefore experiment with two variants of the enhanced NER model. Features in the OOV class take the same value in either variant (i.e., a word is OOV for MADA if and only if it is OOV for BAMA). The difference is in the definition of GlossCap features:

- *BAMA only*. A GlossCap feature is true if the gloss of any analysis returned by BAMA is capitalized.
- *MADA*. A GlossCap feature is true only if the gloss of the analysis selected by MADA is capitalized.

Our use of BAMA in this approach is equivalent to having access to two large static lexical resources, one listing all known wordforms, the other listing all wordforms that can be construed as proper nouns. Arguably, MADA is providing something subtler and more significant. The experiments below test this perspective.

4.2. Experiments

We divided the 221 documents randomly into five partitions, and conducted 5-fold cross-validation, for each fold training a model on the out-of-fold documents and testing its performance on documents in the fold. Performance numbers were then calculated on the combined data (i.e., performance metrics are micro-averaged).

Our performance metrics are precision, recall, and F1, defined over name spans. In order for a model prediction to be counted as correct, both the boundaries and label of a proposed span must match precisely those of a reference span. All other mismatches are failures either of precision or of recall. Precision is the number of predicted mentions whose label and span agreed with a reference mention, divided by the total number of predicted mentions. Recall is the fraction of reference mentions correctly identified by a model. F1 is the harmonic mean of precision and recall.

Table 1 compares the baseline system with the two morphologically enhanced systems described above. It will be observed that while features based on the BAMA analysis afford a marginal improvement over the baseline, the disambiguated features based on the MADA analysis support a much larger improvement. The baseline system is

	Base		MADA	
	S&T	S	S&T	T
F1	0.650	0.695	0.696	0.757
Precision	0.703	0.752	0.723	0.787
Recall	0.604	0.646	0.670	0.730

Table 3: Summary F1, precision, and recall for the baseline BASE and the MADA-enhanced system, with an evaluation on the ACE05 corpus using the standard metric which requires a correct span and tag (S&T), as well as a relaxed metric which requires only a correct span (S)

clearly recall-limited. Introduction of the MADA-based features supports an increase in recall of 7% while simultaneously benefiting precision. Table 2 shows some examples in which the use of MADA improved recall, precision, or both.

The evidence clearly supports the idea that the lexical information provided by the morphological tagger benefits NER. Most of this benefit is attributable to the disambiguation afforded by MADA. In the absence of such disambiguation, it is doubtful whether observed improvements are significant.

5. Error Analysis

In an effort to understand the observed performance, we performed a sequence of error analyses. The first issue we investigated is whether the errors were primarily related to spans or to labels (recall that a correct NER detection requires both a correct span and a correct label). We extended the evaluation to include a second metric which measures only the spans and disregards the labels. The results for the baseline system BASE and the system including MADA are shown in Table 3. The results for the standard metric, marked "S&T", are lower than shown in Table 1 because we are now evaluating on a subset of the ACE 2005 corpus, which appears to be harder. The relative results, however, are quite similar to the previous ones (obtained through cross-validation). From looking at the columns labeled "S", which show the performance evaluating the span match only (and not the tag accuracy), we see that performance improved, but not by much: the baseline system reduced the F1 error only by 13%, while the MADA-enhanced system reduced the F1 error by 20%. We conclude that the harder problem in NER is the correct identification of the spans. We also note that MADA seems to help relatively more when we only evaluate on the spans: the F1 error reduction is 20.5% over the baseline, while in the case of the stricter evaluation, we only reduce F1 error by 13.1%. This makes sense as we expect MADA to help specifically with the identification of constituent boundaries. The analysis suggests that improved morpho-syntactic processing (perhaps also using parsing) may be one of the most fruitful avenues for future work.

To further analyze this issue, we performed a more detailed error analysis on a smaller subset of the ACE 2005 newswire corpus (about 14,000 words). We automatically classified the errors as the following types: recall error (an NE was entirely missed), precision error (an NE was falsely proposed), span error (the proposed span for the NE is wrong), and label error (the exact span for the NE was found but with the wrong label). Note that the 'category "span error" typically includes what would in a simpler classification be called a recall error (our proposed span overlaps with the correct span but we do not identify the correct span) and a precision error (our proposed span is wrong). Furthermore, we are including in this count cases in which we propose several spans that overlap with one gold span. Thus, the numbers in this analysis do not add up in a straightforward manner to the numbers in the previous analysis. We found that 44% of the errors are recall errors (we miss an NE), 16% are precision errors (we propose a false NE), 25% span errors (we propose one or more false span that overlap(s) with a gold span), and only 15% are label errors (the span is correct, the label is not). These numbers confirm the previous result that most errors are related to the span, not the label. Furthermore, the bulk of the span errors are recall errors, as opposed to precision errors, or errors in the exact boundaries of the span: we miss NEs outright.

We investigated the error types in more detail. Surprisingly, most of the recall errors involve common NEs (such as regions, continents , states, world-famous leaders, and generic entity names such as *ministry of health* or *the opposition*), especially geo-political entities. This suggests that the use of a gazetteer could improve the performance of our system. Geo-political entities also are the largest category of precision errors. The most common span error (nearly half the cases) is that the MADA-based system predicts a smaller span than the gold standard. This is consistent with our system having mainly recall errors: spans are not found or are not fully found. This suggests that future work should concentrate on finding conditions on identifying spans and extending the currently found spans. Most label errors (82%) involve geo-political entities, which again suggests further investigation into this type of NE.

We then turned to the question of how different the three systems (baseline, BAMA, MADA) are. Evaluating their predictions against each other, we get F1-measures ranging between 76.5% and 81.7%, suggesting that there is considerable non-overlap in the predictions. This is borne out in further investigation. We first performed an oracle experiment, in which we chose the output of the system which got the right answer (both span and label). This raises the F-measure further to 79.8% (from 71.5% for the MADA system), strongly suggesting that system combination is a promising avenue of future research. We performed some initial system combination experiments. We discuss two systems briefly. If any system suggested a NE, we considered it an NE (**Union**). Note that **Union** does not resolve overlaps, so that it may propose incompatible NE analyses. In **Intersection**, we only propose a NE if all three systems agree on the NE. As expected, we see a recall-precision trade-off, with the **Union** combination achieving a very high recall (73.0%), and the **Intersection** achieving a very high precision (84.3%). Neither of the systems achieves a higher F-measure than our MADA-based system, but we think that further work on system combination is warranted. (We also tried a majority-vote system,

	واصيب منيب ابو منشار (١٩ عاما) برصاصة في قلبه خلال مواجهات بين راشقي الحجارة الفلسطينيين و جنود اسرائيليين ...
BASE	wASyb ⟨PER⟩mnyb Abw mnšAr⟨/PER⟩ (19 ςAmA) brSASħ fy ⟨GPE⟩qlbh⟨/GPE⟩ xlAl mwAjhAt byn rAšqy AlHjArħ ⟨GPE⟩AlflsTynyyn⟨/GPE⟩ wjnwd **AsrAŷylyyn** ...
MADA	wASyb ⟨PER⟩mnyb Abw mnšAr⟨/PER⟩ (19 ςAmA) brSASħ fy qlbh xlAl mwAjhAt byn rAšqy AlHjArħ ⟨GPE⟩AlflsTynyyn⟨/GPE⟩ wjnwd ⟨GPE⟩**AsrAŷylyyn**⟨/GPE⟩ ... Munib Abu Munshar (19 years) was hit with a bullet in the heart during confrontations between Palestinian stone throwers and **Israeli** soldiers ...
	لم تسجل مصافحة بين عرفات وباراك في الجلسة الافتتاحية ...
BASE	lm tsjl mSAfHħ byn ⟨PER⟩ςrfAt⟨/PER⟩ **wbArAk** fy Aljlsħ AlAfttAHyħ ...
MADA	lm tsjl mSAfHħ byn ⟨PER⟩ςrfAt⟨/PER⟩ ⟨PER⟩**wbArAk**⟨/PER⟩ fy Aljlsħ AlAfttAHyħ ... The handshake between Arafat **and Barak** was not recorded in the opening session ...
	... لما كان يطالب به مقدم الشكوى بوب جونز ...
BASE	... lmA kAn yTAlb bh mqdm Alškwý **bwb jwnz** ...
MADA	... lmA kAn yTAlb bh mqdm Alškwý ⟨PER⟩**bwb jwnz**⟨/PER⟩ ... ... what was requested by the plaintiff **Bob Jones** ...
	... ان معارك وقعت بين حركة طالبان والمعارضة الافغانية ...
BASE	... An mςArk wqςt byn ⟨ORG⟩Hrkħ TAlbAn wAlmςArDħ⟨/ORG⟩ **AlAfγAnyħ** ...
MADA	... An mςArk wqςt byn ⟨ORG⟩Hrkħ TAlbAn⟨/ORG⟩ wAlmςArDħ ⟨GPE⟩**AlAfγAnyħ**⟨/GPE⟩ ... ... that battles took place between the Taliban movement and the **Afghan** opposition ...
	... وجرح اسرائيليان اخران خلال هجوم ...
BASE	... wjrH **AsrAŷylyAn** AxrAn xlAl hjwm ...
MADA	... wjrH ⟨GPE⟩**AsrAŷylyAn**⟨/GPE⟩ AxrAn xlAl hjwm ... ... and **two** other **Israelis** were wounded during an attack
	... حركة فتح وحركتا القاومة الاسلامية - حماس - والجهاد ...
BASE	... ⟨ORG⟩Hrkħ ftH wHrktA AlmqAwmħ⟨/ORG⟩ AlAslAmyħ - ⟨ORG⟩HmAs⟨/ORG⟩ - wAljhAd ...
MADA	... ⟨ORG⟩Hrkħ ftH⟨/ORG⟩ ⟨ORG⟩wHrktA AlmqAwmħ AlAslAmyħ -⟨/ORG⟩ ⟨ORG⟩HmAs⟨/ORG⟩ - wAljhAd ... ... Fateh organization and the two Islamic resistance organizations - Hamas and Aljihad ...

Table 2: Examples in which the use of MADA improves over the baseline BASE; underlined words indicate words falsely marked as an NER by BASE, while bold-faced words are NERs which were missed by BASE

which also did not perform better than the MADA-based system.) Furthermore, for applications in which either recall or precision is paramount, our two described systems may already be useful as described.

6. Discussion and Future Work

The research reported in this paper has shown that using a morphological tagger can help in named-entity recognition for Arabic. Our error analysis shows that the major problem in Arabic NER is the proper recognition of the spans, and this is an area where morphological tagging helps. One area we intend to explore is to exploit lemmatization more, which MADA provides, both for the word-specific features, and for the clustering. Lemmatization is a way of reducing lexical data sparseness. Furthermore, the error analysis suggests that future work should concentrate on improved span recognition, for example by improved morpho-syntactic feedback (using a parser perhaps), and by using better resources such as gazetteers.

Acknowledgments

This work was funded under the DARPA GALE program, contract HR0011-06-C-0023. We thank Mona Diab for helpful conversations.

7. References

Shumeet Baluja, Vibhu Mittal, and Rahul Sukthankar. 1999. Applying machine learning for high performance named-entity extraction. In *Pacific Association for Computational Linguistics*.

D.M. Bikel, S. Miller, R. Schwartz, and R. Weischedel. 1997. Nymble: a high-performance learning name-finder. In *Proc. 5th Conference on Applied Natural Language Processing (ANLP-97)*, April.

A. Borthwick. 1999. *A Maximum Entropy Approach to Named Entity Recognition*. Ph.D. thesis, New York University.

Tim Buckwalter. 2004. Buckwalter Arabic Morphological Analyzer Version 2.0. Linguistic Data Consortium, University of Pennsylvania, 2002. LDC Cat alog No.: LDC2004L02, ISBN 1-58563-324-0.

Michael Collins. 2002. Discriminative training methods for hidden Markov models: theory and experiments with perceptron algorithms. In *Proceedings of EMNLP 2002*.

Dayne Freitag. 2004. Trained named entity recognition using distributional clusters. In *Proceedings of EMNLP 2004*.

Nizar Habash and Owen Rambow. 2005. Arabic Tokeniza-

tion, Part-of-Speech Tagging and Morphological Disambiguation in One Fell Swoop. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 573–580, Ann Arbor, Michigan, June. Association for Computational Linguistics.

Nizar Habash, Abdelhadi Soudi, and Tim Buckwalter. 2007. On Arabic Transliteration. In A. van den Bosch and A. Soudi, editors, *Arabic Computational Morphology: Knowledge-based and Empirical Methods*. Springer.

Nizar Habash. 2007. Arabic Morphological Representations for Machine Translation. In A. van den Bosch and A. Soudi, editors, *Arabic Computational Morphology: Knowledge-based and Empirical Methods*. Springer.

H. Isozaki and H. Kazawa. 2002. Efficient support vector classifiers for named entity recognition. In *Proceedings of COLING 2002*.

Mohamed Maamouri, Ann Bies, and Tim Buckwalter. 2004. The Penn Arabic Treebank: Building a large-scale annotated Arabic corpus. In *NEMLAR Conference on Arabic Language Resources and Tools*, Cairo, Egypt.

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

S. Miller, J. Guinness, and A. Zamanian. 2004. Name tagging with word clusters and discriminative training. In *Proceedings of HLT/NAACL 04*.