Recall-Oriented Learning of Named Entities in Arabic Wikipedia

Behrang Mohit* **Nathan Schneider†** **Rishav Bhowmick*** **Kemal Oflazer*** **Noah A. Smith†**
School of Computer Science, Carnegie Mellon University
*P.O. Box 24866, Doha, Qatar †Pittsburgh, PA 15213, USA
{behrang@,nschneid@cs.,rishavb@qatar.,ko@cs.,nasmith@cs.}cmu.edu

Abstract

We consider the problem of NER in Arabic Wikipedia, a semisupervised domain adaptation setting for which we have no labeled training data in the target domain. To facilitate evaluation, we obtain annotations for articles in four topical groups, allowing annotators to identify domain-specific entity types in addition to standard categories. Standard supervised learning on newswire text leads to poor target-domain recall. We train a sequence model and show that a simple modification to the online learner—a loss function encouraging it to "arrogantly" favor recall over precision—substantially improves recall and F_1. We then adapt our model with self-training on unlabeled target-domain data; enforcing the same recall-oriented bias in the self-training stage yields marginal gains.[1]

1 Introduction

This paper considers named entity recognition (NER) in text that is different from most past research on NER. Specifically, we consider Arabic Wikipedia articles with diverse topics beyond the commonly-used news domain. These data challenge past approaches in two ways:

First, Arabic is a morphologically rich language (Habash, 2010). Named entities are referenced using complex syntactic constructions (cf. English NEs, which are primarily sequences of proper nouns). The Arabic script suppresses most vowels, increasing lexical ambiguity, and lacks capitalization, a key clue for English NER.

Second, much research has focused on the use of *news* text for system building and evaluation. Wikipedia articles are not news, belonging instead to a wide range of domains that are not clearly delineated. One hallmark of this divergence between Wikipedia and the news domain is a difference in the distributions of named entities. Indeed, the classic named entity types (person, organization, location) may not be the most apt for articles in other domains (e.g., scientific or social topics). On the other hand, Wikipedia is a large dataset, inviting semisupervised approaches.

In this paper, we describe advances on the problem of NER in Arabic Wikipedia. The techniques are general and make use of well-understood building blocks. Our contributions are:

- A small corpus of articles annotated in a new scheme that provides more freedom for annotators to adapt NE analysis to new domains;
- An "arrogant" learning approach designed to boost recall in supervised training as well as self-training; and
- An empirical evaluation of this technique as applied to a well-established discriminative NER model and feature set.

Experiments show consistent gains on the challenging problem of identifying named entities in Arabic Wikipedia text.

2 Arabic Wikipedia NE Annotation

Most of the effort in NER has been focused around a small set of domains and general-purpose entity classes relevant to those domains—especially the categories PER(SON), ORG(ANIZATION), and LOC(ATION) (POL), which are highly prominent in news text. Arabic is no exception: the publicly available NER corpora—ACE (Walker et al., 2006), ANER (Benajiba et al., 2008), and OntoNotes (Hovy et al., 2006)—all are in the news domain.[2] However,

[1] The annotated dataset and a supplementary document with additional details of this work can be found at: http://www.ark.cs.cmu.edu/AQMAR

[2] OntoNotes contains news-related text. ACE includes some text from blogs. In addition to the POL classes, both corpora include additional NE classes such as facility, event, product, vehicle, etc. These entities are infrequent and may not be comprehensive enough to cover the larger set of pos-

Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, pages 162–173, Avignon, France, April 23 - 27 2012. ©2012 Association for Computational Linguistics

	History	Science	Sports	Technology
dev:	Damascus	Atom	Raúl Gonzáles	Linux
	Imam Hussein Shrine	Nuclear power	Real Madrid	Solaris
test:	Crusades	Enrico Fermi	2004 Summer Olympics	Computer
	Islamic Golden Age	Light	Christiano Ronaldo	Computer Software
	Islamic History	Periodic Table	Football	Internet
	Ibn Tolun Mosque	Physics	Portugal football team	Richard Stallman
	Ummaya Mosque	Muhammad al-Razi	FIFA World Cup	X Window System

Table 1: Translated titles of Arabic Wikipedia articles in our development and test sets, and some NEs with standard and article-specific classes. Additionally, Prussia and Amman were reserved for training annotators, and Gulf War for estimating inter-annotator agreement.

Claudio Filippone (PER) كلوديو فيلبون; Linux (SOFTWARE) لينكس; Spanish League (CHAMPIONSHIPS) الدوري الاسباني; proton (PARTICLE) بروتون; nuclear radiation (GENERIC-MISC) الاشعاع النووي; Real Zaragoza (ORG) ريال سرقسطة

appropriate entity classes will vary widely by domain; occurrence rates for entity classes are quite different in news text vs. Wikipedia, for instance (Balasuriya et al., 2009). This is abundantly clear in technical and scientific discourse, where much of the terminology is domain-specific, but it holds elsewhere. Non-POL entities in the history domain, for instance, include important events (wars, famines) and cultural movements (*romanticism*). Ignoring such domain-critical entities likely limits the usefulness of the NE analysis.

Recognizing this limitation, some work on NER has sought to codify more robust inventories of general-purpose entity types (Sekine et al., 2002; Weischedel and Brunstein, 2005; Grouin et al., 2011) or to enumerate domain-specific types (Settles, 2004; Yao et al., 2003). Coarse, general-purpose categories have also been used for semantic tagging of nouns and verbs (Ciaramita and Johnson, 2003). Yet as the number of classes or domains grows, rigorously documenting and organizing the classes—even for a single language—requires intensive effort. Ideally, an NER system would refine the traditional classes (Hovy et al., 2011) or identify new entity classes when they arise in new domains, adapting to new data. For this reason, we believe it is valuable to consider NER systems that identify (but do not necessarily label) entity mentions, and also to consider annotation schemes that allow annotators more freedom in defining entity classes.

Our aim in creating an annotated dataset is to provide a testbed for *evaluation* of new NER models. We will use these data as development and testing examples, but not as training data. In §4 we will discuss our semisupervised approach to learning, which leverages ACE and ANER data as an annotated training corpus.

2.1 Annotation Strategy

We conducted a small annotation project on Arabic Wikipedia articles. Two college-educated native Arabic speakers annotated about 3,000 sentences from 31 articles. We identified four topical areas of interest—history, technology, science, and sports—and browsed these topics until we had found 31 articles that we deemed satisfactory on the basis of length (at least 1,000 words), cross-lingual linkages (associated articles in English, German, and Chinese[3]), and subjective judgments of quality. The list of these articles along with sample NEs are presented in table 1. These articles were then preprocessed to extract main article text (eliminating tables, lists, info-boxes, captions, etc.) for annotation.

Our approach follows ACE guidelines (LDC, 2005) in identifying NE boundaries and choosing POL tags. In addition to this traditional form of annotation, annotators were encouraged to articulate one to three *salient, article-specific* entity categories per article. For example, names of particles (e.g., *proton*) are highly salient in the Atom article. Annotators were asked to read the entire article first, and then to decide which non-traditional classes of entities would be important in the context of article. In some cases, annotators reported using heuristics (such as being proper

sible NEs (Sekine et al., 2002). Nezda et al. (2006) annotated and evaluated an Arabic NE corpus with an extended set of 18 classes (including temporal and numeric entities); this corpus has not been released publicly.

[3]These three languages have the most articles on Wikipedia. Associated articles here are those that have been manually hyperlinked from the Arabic page as cross-lingual correspondences. They are not translations, but if the associations are accurate, these articles should be topically similar to the Arabic page that links to them.

Token position agreement rate	92.6%	Cohen's κ: 0.86
Token agreement rate	88.3%	Cohen's κ: 0.86
Token F_1 between annotators	91.0%	
Entity boundary match F_1	94.0%	
Entity category match F_1	87.4%	

Table 2: Inter-annotator agreement measurements.

nouns or having an English translation which is conventionally capitalized) to help guide their determination of non-canonical entities and entity classes. Annotators produced written descriptions of their classes, including example instances.

This scheme was chosen for its flexibility: in contrast to a scenario with a fixed ontology, annotators required minimal training beyond the POL conventions, and did not have to worry about delineating custom categories precisely enough that they would extend straightforwardly to other topics or domains. Of course, we expect inter-annotator variability to be greater for these open-ended classification criteria.

2.2 Annotation Quality Evaluation

During annotation, two articles (Prussia and Amman) were reserved for training annotators on the task. Once they were accustomed to annotation, both independently annotated a third article. We used this 4,750-word article (Gulf War, حرب الخليج الثانية) to measure inter-annotator agreement. Table 2 provides scores for token-level agreement measures and entity-level F_1 between the two annotated versions of the article.[4]

These measures indicate strong agreement for locating and categorizing NEs both at the token and chunk levels. Closer examination of agreement scores shows that PER and MIS classes have the lowest rates of agreement. That the miscellaneous class, used for infrequent or article-specific NEs, receives poor agreement is unsurprising. The low agreement on the PER class seems to be due to the use of titles and descriptive terms in personal names. Despite explicit guidelines to exclude the titles, annotators disagreed on the inclusion of descriptors that disambiguate the NE (e.g., *the father* in جرج بوش الآب: George Bush, the father).

[4]The position and boundary measures ignore the distinctions between the POLM classes. To avoid artificial inflation of the token and token position agreement rates, we exclude the 81% of tokens tagged by both annotators as not belonging to an entity.

History: Gulf War, Prussia, Damascus, Crusades			
		WAR_CONFLICT	● ● ●
Science: Atom, Periodic table			
	THEORY ●	CHEMICAL	● ●
	NAME_ROMAN ●	PARTICLE	● ●
Sports: Football, Raúl Gonzáles			
	SPORT ○	CHAMPIONSHIP	●
	AWARD ○	NAME_ROMAN	●
Technology: Computer, Richard Stallman			
	COMPUTER_VARIETY ○	SOFTWARE	●
		COMPONENT	●

Table 3: Custom NE categories suggested by one or both annotators for 10 articles. Article titles are translated from Arabic. ● indicates that both annotators volunteered a category for an article; ○ indicates that only one annotator suggested the category. Annotators were not given a predetermined set of possible categories; rather, category matches between annotators were determined by post hoc analysis. NAME_ROMAN indicates an NE rendered in Roman characters.

2.3 Validating Category Intuitions

To investigate the variability between annotators with respect to custom category intuitions, we asked our two annotators to independently read 10 of the articles in the data (scattered across our four focus domains) and suggest up to 3 custom categories for each. We assigned short names to these suggestions, seen in table 3. In 13 cases, both annotators suggested a category for an article that was essentially the same (●); three such categories spanned multiple articles. In three cases a category was suggested by only one annotator (○).[5] Thus, we see that our annotators were generally, but not entirely, consistent with each other in their creation of custom categories. Further, almost all of our article-specific categories correspond to classes in the extended NE taxonomy of (Sekine et al., 2002), which speaks to the reasonableness of both sets of categories—and by extension, our open-ended annotation process.

Our annotation of named entities outside of the traditional POL classes creates a useful resource for entity detection and recognition in new domains. Even the ability to detect non-canonical types of NEs should help applications such as QA and MT (Toral et al., 2005; Babych and Hartley, 2003). Possible avenues for future work include annotating and projecting non-canonical

[5]When it came to tagging NEs, one of the two annotators was assigned to each article. Custom categories only suggested by the other annotator were ignored.

NEs from English articles to their Arabic counterparts (Hassan et al., 2007), automatically clustering non-canonical types of entities into article-specific or cross-article classes (cf. Frietag, 2004), or using non-canonical classes to improve the (author-specified) article categories in Wikipedia.

Hereafter, we merge all article-specific categories with the generic MIS category. The proportion of entity mentions that are tagged as MIS, while varying to a large extent by document, is a major indication of the gulf between the news data (<10%) and the Wikipedia data (53% for the development set, 37% for the test set).

Below, we aim to develop entity detection models that generalize beyond the traditional POL entities. We do not address here the challenges of automatically *classifying* entities or inferring non-canonical groupings.

3 Data

Table 4 summarizes the various corpora used in this work.[6] Our NE-annotated Wikipedia subcorpus, described above, consists of several Arabic Wikipedia articles from four focus domains.[7] We do not use these for supervised training data; they serve only as development and test data. A larger set of Arabic Wikipedia articles, selected on the basis of quality heuristics, serves as unlabeled data for semisupervised learning.

Our out-of-domain labeled NE data is drawn from the ANER (Benajiba et al., 2007) and ACE-2005 (Walker et al., 2006) newswire corpora. Entity types in this data are POL categories (PER, ORG, LOC) and MIS. Portions of the ACE corpus were held out as development and test data; the remainder is used in training.

4 Models

Our starting point for statistical NER is a feature-based linear model over sequences, trained using the structured perceptron (Collins, 2002).[8]

In addition to lexical and morphological[9] features known to work well for Arabic NER (Benajiba et al., 2008; Abdul-Hamid and Darwish, 2010), we incorporate some additional features enabled by Wikipedia. We do not employ a gazetteer, as the construction of a broad-domain gazetteer is a significant undertaking orthogonal to the challenges of a new text domain like Wikipedia.[10] A descriptive list of our features is available in the supplementary document.

Training	*words*	*NEs*
ACE+ANER	212,839	15,796
Wikipedia (unlabeled, 397 docs)	1,110,546	—
Development		
ACE	7,776	638
Wikipedia (4 domains, 8 docs)	21,203	2,073
Test		
ACE	7,789	621
Wikipedia (4 domains, 20 docs)	52,650	3,781

Table 4: Number of words (entity mentions) in data sets.

We use a first-order structured perceptron; none of our features consider more than a pair of consecutive BIO labels at a time. The model enforces the constraint that NE sequences must begin with B (so the bigram $\langle O, I\rangle$ is disallowed).

Training this model on ACE and ANER data achieves performance comparable to the state of the art (F_1-measure[11] above 69%), but fares much worse on our Wikipedia test set (F_1-measure around 47%); details are given in §5.

4.1 Recall-Oriented Perceptron

By augmenting the perceptron's online update with a *cost* function term, we can incorporate a task-dependent notion of error into the objective, as with structured SVMs (Taskar et al., 2004; Tsochantaridis et al., 2005). Let $c(\boldsymbol{y}, \boldsymbol{y}')$ denote a measure of error when $\boldsymbol{y}$ is the correct label sequence but $\boldsymbol{y}'$ is predicted. For observed sequence $\boldsymbol{x}$ and feature weights (model parameters) $\mathbf{w}$, the structured hinge loss is $\ell_{hinge}(\boldsymbol{x}, \boldsymbol{y}, \mathbf{w}) =$

$$\max_{\boldsymbol{y}'} \left(\mathbf{w}^\top \mathbf{g}(\boldsymbol{x}, \boldsymbol{y}') + c(\boldsymbol{y}, \boldsymbol{y}')\right) - \mathbf{w}^\top \mathbf{g}(\boldsymbol{x}, \boldsymbol{y}) \quad (1)$$

The maximization problem inside the parentheses is known as *cost-augmented decoding*. If c fac-

[6] Additional details appear in the supplement.

[7] We downloaded a snapshot of Arabic Wikipedia (`http://ar.wikipedia.org`) on 8/29/2009 and preprocessed the articles to extract main body text and metadata using the `mwlib` package for Python (PediaPress, 2010).

[8] A more leisurely discussion of the structured perceptron and its connection to empirical risk minimization can be found in the supplementary document.

[9] We obtain morphological analyses from the MADA tool (Habash and Rambow, 2005; Roth et al., 2008).

[10] A gazetteer ought to yield further improvements in line with previous findings in NER (Ratinov and Roth, 2009).

[11] Though optimizing NER systems for F_1 has been called into question (Manning, 2006), no alternative metric has achieved widespread acceptance in the community.

tors similarly to the feature function $\mathbf{g}(\boldsymbol{x}, \boldsymbol{y})$, then we can increase penalties for $\boldsymbol{y}$ that have more local mistakes. This raises the learner's awareness about how it will be evaluated. Incorporating cost-augmented decoding into the perceptron leads to this decoding step:

$$\hat{\boldsymbol{y}} \leftarrow \arg\max_{\boldsymbol{y}'} \left(\mathbf{w}^\top \mathbf{g}(\boldsymbol{x}, \boldsymbol{y}') + c(\boldsymbol{y}, \boldsymbol{y}') \right), \quad (2)$$

which amounts to performing stochastic subgradient ascent on an objective function with the Eq. 1 loss (Ratliff et al., 2006).

In this framework, cost functions can be formulated to distinguish between different types of errors made during training. For a tag sequence $\boldsymbol{y} = \langle y_1, y_2, \ldots, y_M \rangle$, Gimpel and Smith (2010b) define word-local cost functions that differently penalize precision errors (i.e., $y_i = O \wedge \hat{y}_i \neq O$ for the ith word), recall errors ($y_i \neq O \wedge \hat{y}_i = O$), and entity class/position errors (other cases where $y_i \neq \hat{y}_i$). As will be shown below, a key problem in cross-domain NER is poor *recall*, so we will penalize recall errors more severely:

$$c(\boldsymbol{y}, \boldsymbol{y}') = \sum_{i=1}^{M} \begin{cases} 0 & \text{if } y_i = y'_i \\ \beta & \text{if } y_i \neq O \wedge y'_i = O \\ 1 & \text{otherwise} \end{cases} \quad (3)$$

for a penalty parameter $\beta > 1$. We call our learner the "recall-oriented" perceptron (ROP).

We note that Minkov et al. (2006) similarly explored the recall vs. precision tradeoff in NER. Their technique was to directly tune the weight of a single feature—the feature marking O (non-entity tokens); a lower weight for this feature will incur a greater penalty for predicting O. Below we demonstrate that our method, which is less coarse, is more successful in our setting.[12]

In our experiments we will show that injecting "arrogance" into the learner via the recall-oriented loss function substantially improves recall, especially for non-POL entities (§5.3).

4.2 Self-Training and Semisupervised Learning

As we will show experimentally, the differences between news text and Wikipedia text call for domain adaptation. In the case of Arabic Wikipedia, there is no available labeled training data. Yet the available *unlabeled* data is vast, so we turn to semisupervised learning.

Input: labeled data $\langle \langle \boldsymbol{x}^{(n)}, \boldsymbol{y}^{(n)} \rangle \rangle_{n=1}^{N}$; unlabeled data $\langle \bar{\boldsymbol{x}}^{(j)} \rangle_{j=1}^{J}$; supervised learner L; number of iterations T'
Output: $\mathbf{w}$
$\mathbf{w} \leftarrow L(\langle \langle \boldsymbol{x}^{(n)}, \boldsymbol{y}^{(n)} \rangle \rangle_{n=1}^{N})$
for $t = 1$ **to** T' **do**
  for $j = 1$ **to** J **do**
    $\hat{\boldsymbol{y}}^{(j)} \leftarrow \arg\max_{\boldsymbol{y}} \mathbf{w}^\top \mathbf{g}(\bar{\boldsymbol{x}}^{(j)}, \boldsymbol{y})$
  $\mathbf{w} \leftarrow L(\langle \langle \boldsymbol{x}^{(n)}, \boldsymbol{y}^{(n)} \rangle \rangle_{n=1}^{N} \cup \langle \langle \bar{\boldsymbol{x}}^{(j)}, \hat{\boldsymbol{y}}^{(j)} \rangle \rangle_{j=1}^{J})$

Algorithm 1: Self-training.

Here we adapt self-training, a simple technique that leverages a supervised learner (like the perceptron) to perform semisupervised learning (Clark et al., 2003; Mihalcea, 2004; McClosky et al., 2006). In our version, a model is trained on the labeled data, then used to label the unlabeled target data. We iterate between training on the hypothetically-labeled target data plus the original labeled set, and relabeling the target data; see Algorithm 1. Before self-training, we remove sentences hypothesized not to contain any named entity mentions, which we found avoids further encouragement of the model toward low recall.

5 Experiments

We investigate two questions in the context of NER for Arabic Wikipedia:

- **Loss function:** Does integrating a cost function into our learning algorithm, as we have done in the **recall-oriented perceptron** (§4.1), improve recall and overall performance on Wikipedia data?
- **Semisupervised learning for domain adaptation:** Can our models benefit from large amounts of unlabeled Wikipedia data, in addition to the (out-of-domain) labeled data? We experiment with a self-training phase following the fully supervised learning phase.

We report experiments for the possible combinations of the above ideas. These are summarized in table 5. Note that the recall-oriented perceptron can be used for the supervised learning phase, for the self-training phase, or both. This leaves us with the following combinations:

- **reg/none** (baseline): regular supervised learner.
- **ROP/none**: recall-oriented supervised learner.

[12] The distinction between the techniques is that our cost function adjusts the *whole* model in order to perform better at recall on the training data.

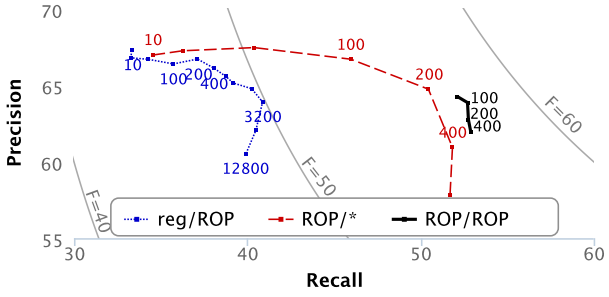


Figure 1: Tuning the recall-oriented cost parameter for different learning settings. We optimized for development set F_1, choosing penalty $\beta = 200$ for recall-oriented supervised learning (in the plot, ROP/*—this is regardless of whether a stage of self-training will follow); $\beta = 100$ for recall-oriented self-training following recall-oriented supervised learning (ROP/ROP); and $\beta = 3200$ for recall-oriented self-training following regular supervised learning (reg/ROP).

- **reg/reg**: standard self-training setup.
- **ROP/reg**: recall-oriented supervised learner, followed by standard self-training.
- **reg/ROP**: regular supervised model as the initial labeler for recall-oriented self-training.
- **ROP/ROP** (the "double ROP" condition): recall-oriented supervised model as the initial labeler for recall-oriented self-training. Note that the two ROPs can use different cost parameters.

For evaluating our models we consider the named entity *detection* task, i.e., recognizing which spans of words constitute entities. This is measured by per-entity precision, recall, and F_1.[13] To measure statistical significance of differences between models we use Gimpel and Smith's (2010) implementation of the paired bootstrap resampler of (Koehn, 2004), taking 10,000 samples for each comparison.

5.1 Baseline

Our baseline is the perceptron, trained on the POL entity boundaries in the ACE+ANER corpus (**reg/none**).[14] Development data was used to select the number of iterations (10). We performed 3-fold cross-validation on the ACE data and found wide variance in the *in-domain* entity detection performance of this model:

	P	**R**	**F$_1$**
fold 1	70.43	63.08	66.55
fold 2	87.48	81.13	84.18
fold 3	65.09	51.13	57.27
average	74.33	65.11	69.33

(Fold 1 corresponds to the ACE test set described in table 4.) We also trained the model to perform POL detection and classification, achieving nearly identical results in the 3-way cross-validation of ACE data. From these data we conclude that our baseline is on par with the state of the art for Arabic NER on ACE news text (Abdul-Hamid and Darwish, 2010).[15]

Here is the performance of the baseline entity detection model on our 20-article test set:[16]

	P	**R**	**F$_1$**
technology	60.42	20.26	30.35
science	64.96	25.73	36.86
history	63.09	35.58	45.50
sports	71.66	59.94	65.28
overall	66.30	35.91	46.59

Unsurprisingly, performance on Wikipedia data varies widely across article domains and is much lower than in-domain performance. Precision scores fall between 60% and 72% for all domains, but recall in most cases is far worse. Miscellaneous class recall, in particular, suffers badly (under 10%)—which partially accounts for the poor recall in science and technology articles (they have by far the highest proportion of MIS entities).

5.2 Self-Training

Following Clark et al. (2003), we applied self-training as described in Algorithm 1, with the perceptron as the supervised learner. Our unlabeled data consists of 397 Arabic Wikipedia articles (1 million words) selected at random from all articles exceeding a simple length threshold (1,000 words); see table 4. We used only one iteration ($T' = 1$), as experiments on development data showed no benefit from additional rounds. Several rounds of self-training hurt performance,

[13] Only entity spans that exactly match the gold spans are counted as correct. We calculated these scores with the `conlleval.pl` script from the CoNLL 2003 shared task.

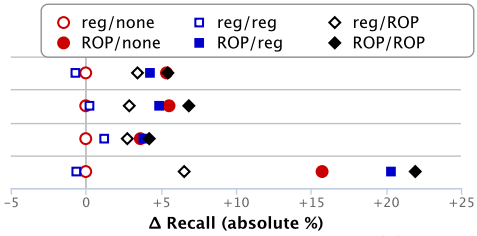
[14] In keeping with prior work, we ignore non-POL categories for the ACE evaluation.

[15] Abdul-Hamid and Darwish report as their best result a macroaveraged F_1-score of 76. As they do not specify which data they used for their held-out test set, we cannot perform a direct comparison. However, our feature set is nearly a superset of their best feature set, and their result lies well within the range of results seen in our cross-validation folds.

[16] Our Wikipedia evaluations use models trained on POLM entity boundaries in ACE. Per-domain and overall scores are microaverages across articles.

Supervised	Self-training: none			Self-training: reg			Self-training: ROP		
reg	66.3	35.9	46.59	66.7	35.6	46.41	59.2	40.3	47.97
ROP	60.9	44.7	**51.59**	59.8	46.2	**52.11**	58.0	47.4	**52.16**

Table 5: Entity detection precision, recall, and F_1 for each learning setting, microaveraged across the 24 articles in our Wikipedia test set. Rows differ in the supervised learning condition on the ACE+ANER data (regular vs. recall-oriented perceptron). Columns indicate whether this supervised learning phase was followed by self-training on unlabeled Wikipedia data, and if so which version of the perceptron was used for self-training.



	entities	words	baseline recall
PER	1081	1743	49.95
ORG	286	637	23.92
LOC	1019	1413	61.43
MIS	1395	2176	9.30
overall	3781	5969	35.91

Figure 2: Recall improvement over baseline in the test set by gold NER category, counts for those categories in the data, and recall scores for our baseline model. Markers in the plot indicate different experimental settings corresponding to cells in table 5.

an effect attested in earlier research (Curran et al., 2007) and sometimes known as "semantic drift."

Results are shown in table 5. We find that standard self-training (the middle column) has very little impact on performance.[17] Why is this the case? We venture that poor baseline recall and the domain variability *within* Wikipedia are to blame.

5.3 Recall-Oriented Learning

The recall-oriented bias can be introduced in either or both of the stages of our semisupervised learning framework: in the supervised learning phase, modifying the objective of our baseline (§5.1); and within the self-training algorithm (§5.2).[18] As noted in §4.1, the aim of this approach is to discourage recall errors (false negatives), which are the chief difficulty for the news text–trained model in the new domain. We selected the value of the false positive penalty for cost-augmented decoding, β, using the development data (figure 1).

The results in table 5 demonstrate improvements due to the recall-oriented bias in both stages of learning.[19] When used in the supervised phase (bottom left cell), the recall gains are substantial—nearly 9% over the baseline. Integrating this bias within self-training (last column of the table) produces a more modest improvement (less than 3%) relative to the baseline. In both cases, the improvements to recall more than compensate for the amount of degradation to precision. This trend is robust: wherever the recall-oriented perceptron is added, we observe improvements in both recall and F_1. Perhaps surprisingly, these gains are somewhat additive: using the ROP in both learning phases gives a small (though not always significant) gain over alternatives (standard supervised perceptron, no self-training, or self-training with a standard perceptron). In fact, when the standard supervised learner is used, recall-oriented self-training succeeds despite the ineffectiveness of standard self-training.

Performance breakdowns by (gold) class, figure 2, and domain, figure 3, further attest to the robustness of the overall results. The most dramatic gains are in miscellaneous class recall—each form of the recall bias produces an improvement, and using this bias in both the supervised and self-training phases is clearly most successful for miscellaneous entities. Correspondingly, the technology and science domains (in which this class dominates—83% and 61% of mentions, ver-

[17] In neither case does regular self-training produce a significantly different F_1 score than no self-training.

[18] Standard Viterbi decoding was used to *label* the data within the self-training algorithm; note that cost-augmented decoding only makes sense in learning, not as a prediction technique, since it deliberately introduces errors relative to a correct output that must be provided.

[19] In terms of F_1, the worst of the 3 models with the ROP supervised learner significantly outperforms the best model with the regular supervised learner ($p < 0.005$). The improvements due to self-training are marginal, however: ROP self-training produces a significant gain only following regular supervised learning ($p < 0.05$).

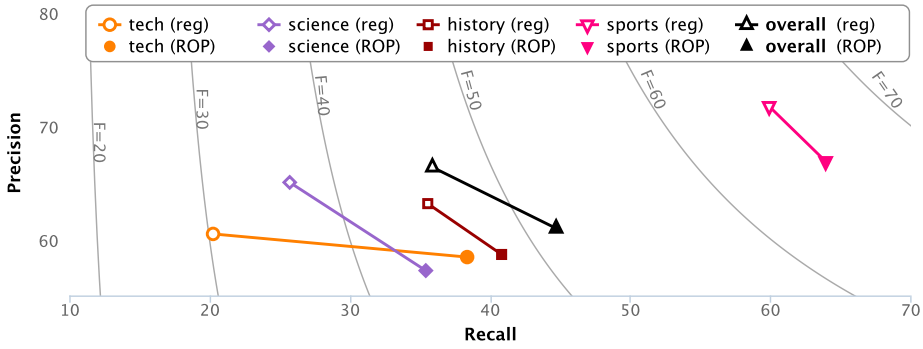


Figure 3: Supervised learner precision vs. recall as evaluated on Wikipedia test data in different topical domains. The regular perceptron (baseline model) is contrasted with ROP. No self-training is applied.

sus 6% and 12% for history and sports, respectively) receive the biggest boost. Still, the gaps between domains are not entirely removed.

Most improvements relate to the reduction of false negatives, which fall into three groups: (a) entities occurring infrequently or partially in the labeled training data (e.g. *uranium*); (b) domain-specific entities sharing lexical or contextual features with the POL entities (e.g. *Linux*, *titanium*); and (c) words with Latin characters, common in the science and technology domains. (a) and (b) are mostly transliterations into Arabic.

An alternative—and simpler—approach to controlling the precision-recall tradeoff is the Minkov et al. (2006) strategy of tuning a single feature weight subsequent to learning (see §4.1 above). We performed an oracle experiment to determine how this compares to recall-oriented learning in our setting. An oracle trained with the method of Minkov et al. outperforms the three models in table 5 that use the *regular* perceptron for the supervised phase of learning, but *underperforms* the supervised ROP conditions.[20]

Overall, we find that incorporating the recall-oriented bias in learning is fruitful for adapting to Wikipedia because the gains in recall outpace the damage to precision.

6 Discussion

To our knowledge, this work is the first suggestion that substantively modifying the *supervised* learning criterion in a resource-rich domain can reap benefits in subsequent *semisupervised* application in a new domain. Past work has looked at regularization (Chelba and Acero, 2006) and feature design (Daumé III, 2007); we alter the loss function. Not surprisingly, the double-ROP approach harms performance on the original domain (on ACE data, we achieve 55.41% F_1, far below the standard perceptron). Yet we observe that models can be prepared for adaptation even before a learner is exposed a new domain, sacrificing performance in the original domain.

The recall-oriented bias is not merely encouraging the learner to identify entities already seen in training. As recall increases, so does the number of new entity types recovered by the model: of the 2,070 NE types in the test data that were never seen in training, only 450 were ever found by the baseline, versus 588 in the reg/ROP condition, 632 in the ROP/none condition, and 717 in the double-ROP condition.

We note finally that our method is a simple extension to the standard structured perceptron; cost-augmented inference is often no more expensive than traditional inference, and the algorithmic change is equivalent to adding one additional feature. Our recall-oriented cost function is parameterized by a single value, β; recall is highly sensitive to the choice of this value (figure 1 shows how we tuned it on development data), and thus we anticipate that, in general, such tuning will be essential to leveraging the benefits of arrogance.

7 Related Work

Our approach draws on insights from work in the areas of NER, domain adaptation, NLP with Wikipedia, and semisupervised learning. As all are broad areas of research, we highlight only the most relevant contributions here.

Research in Arabic NER has been focused on compiling and optimizing the gazetteers and fea-

[20] Tuning the O feature weight to optimize for F_1 on our test set, we found that oracle precision would be 66.2, recall would be 39.0, and F_1 would be 49.1. The F_1 score of our best model is nearly 3 points higher than the Minkov et al.–style oracle, and over 4 points higher than the non-oracle version where the development set is used for tuning.

ture sets for standard sequential modeling algorithms (Benajiba et al., 2008; Farber et al., 2008; Shaalan and Raza, 2008; Abdul-Hamid and Darwish, 2010). We make use of features identified in this prior work to construct a strong baseline system. We are unaware of any Arabic NER work that has addressed diverse text domains like Wikipedia. Both the English and Arabic versions of Wikipedia have been used, however, as resources in service of traditional NER (Kazama and Torisawa, 2007; Benajiba et al., 2008). Attia et al. (2010) heuristically induce a mapping between Arabic Wikipedia and Arabic WordNet to construct Arabic NE gazetteers.

Balasuriya et al. (2009) highlight the substantial divergence between entities appearing in English Wikipedia versus traditional corpora, and the effects of this divergence on NER performance. There is evidence that models trained on Wikipedia data generalize and perform well on corpora with narrower domains. Nothman et al. (2009) and Balasuriya et al. (2009) show that NER models trained on both automatically and manually annotated Wikipedia corpora perform reasonably well on news corpora. The reverse scenario does not hold for models trained on news text, a result we also observe in Arabic NER. Other work has gone beyond the entity detection problem: Florian et al. (2004) additionally predict within-document entity coreference for Arabic, Chinese, and English ACE text, while Cucerzan (2007) aims to resolve every mention detected in English Wikipedia pages to a canonical article devoted to the entity in question.

The domain and topic diversity of NEs has been studied in the framework of domain adaptation research. A group of these methods use self-training and select the most informative features and training instances to adapt a source domain learner to the new target domain. Wu et al. (2009) bootstrap the NER leaner with a subset of unlabeled instances that bridge the source and target domains. Jiang and Zhai (2006) and Daumé III (2007) make use of some labeled target-domain data to tune or augment the features of the source model towards the target domain. Here, in contrast, we use labeled target-domain data only for tuning and evaluation. Another important distinction is that domain variation in this prior work is restricted to topically-related corpora (e.g. newswire vs. broadcast news), whereas in our work, major topical differences distinguish the training and test corpora—and consequently, their salient NE classes. In these respects our NER setting is closer to that of Florian et al. (2010), who recognize English entities in noisy text, (Surdeanu et al., 2011), which concerns information extraction in a topically distinct target domain, and (Dalton et al., 2011), which addresses English NER in noisy and topically divergent text.

Self-training (Clark et al., 2003; Mihalcea, 2004; McClosky et al., 2006) is widely used in NLP and has inspired related techniques that learn from automatically labeled data (Liang et al., 2008; Petrov et al., 2010). Our self-training procedure differs from some others in that we use all of the automatically labeled examples, rather than filtering them based on a confidence score.

Cost functions have been used in non-structured classification settings to penalize certain types of errors more than others (Chan and Stolfo, 1998; Domingos, 1999; Kiddon and Brun, 2011). The goal of optimizing our structured NER model for recall is quite similar to the scenario explored by Minkov et al. (2006), as noted above.

8 Conclusion

We explored the problem of learning an NER model suited to domains for which no labeled training data are available. A loss function to encourage recall over precision during supervised discriminative learning substantially improves recall and overall entity detection performance, especially when combined with a semisupervised learning regimen incorporating the same bias. We have also developed a small corpus of Arabic Wikipedia articles via a flexible entity annotation scheme spanning four topical domains (publicly available at `http://www.ark.cs.cmu.edu/AQMAR`).

Acknowledgments

We thank Mariem Fekih Zguir and Reham Al Tamime for assistance with annotation, Michael Heilman for his tagger implementation, and Nizar Habash and colleagues for the MADA toolkit. We thank members of the ARK group at CMU, Hal Daumé, and anonymous reviewers for their valuable suggestions. This publication was made possible by grant NPRP-08-485-1-083 from the Qatar National Research Fund (a member of the Qatar Foundation). The statements made herein are solely the responsibility of the authors.

References

Ahmed Abdul-Hamid and Kareem Darwish. 2010. Simplified feature set for Arabic named entity recognition. In *Proceedings of the 2010 Named Entities Workshop*, pages 110–115, Uppsala, Sweden, July. Association for Computational Linguistics.

Mohammed Attia, Antonio Toral, Lamia Tounsi, Monica Monachini, and Josef van Genabith. 2010. An automatically built named entity lexicon for Arabic. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta, May. European Language Resources Association (ELRA).

Bogdan Babych and Anthony Hartley. 2003. Improving machine translation quality with automatic named entity recognition. In *Proceedings of the 7th International EAMT Workshop on MT and Other Language Technology Tools*, EAMT '03.

Dominic Balasuriya, Nicky Ringland, Joel Nothman, Tara Murphy, and James R. Curran. 2009. Named entity recognition in Wikipedia. In *Proceedings of the 2009 Workshop on The People's Web Meets NLP: Collaboratively Constructed Semantic Resources*, pages 10–18, Suntec, Singapore, August. Association for Computational Linguistics.

Yassine Benajiba, Paolo Rosso, and José Miguel BenedíRuiz. 2007. ANERsys: an Arabic named entity recognition system based on maximum entropy. In Alexander Gelbukh, editor, *Proceedings of CICLing*, pages 143–153, Mexico City, Mexio. Springer.

Yassine Benajiba, Mona Diab, and Paolo Rosso. 2008. Arabic named entity recognition using optimized feature sets. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 284–293, Honolulu, Hawaii, October. Association for Computational Linguistics.

Philip K. Chan and Salvatore J. Stolfo. 1998. Toward scalable learning with non-uniform class and cost distributions: a case study in credit card fraud detection. In *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining*, pages 164–168, New York City, New York, USA, August. AAAI Press.

Ciprian Chelba and Alex Acero. 2006. Adaptation of maximum entropy capitalizer: Little data can help a lot. *Computer Speech and Language*, 20(4):382–399.

Massimiliano Ciaramita and Mark Johnson. 2003. Supersense tagging of unknown nouns in WordNet. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*, pages 168–175.

Stephen Clark, James Curran, and Miles Osborne. 2003. Bootstrapping POS-taggers using unlabelled data. In Walter Daelemans and Miles Osborne, editors, *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 49–55.

Michael Collins. 2002. Discriminative training methods for hidden Markov models: theory and experiments with perceptron algorithms. In *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1–8, Stroudsburg, PA, USA. Association for Computational Linguistics.

Silviu Cucerzan. 2007. Large-scale named entity disambiguation based on Wikipedia data. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 708–716, Prague, Czech Republic, June.

James R. Curran, Tara Murphy, and Bernhard Scholz. 2007. Minimising semantic drift with Mutual Exclusion Bootstrapping. In *Proceedings of PACLING, 2007*.

Jeffrey Dalton, James Allan, and David A. Smith. 2011. Passage retrieval for incorporating global evidence in sequence labeling. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management (CIKM '11)*, pages 355–364, Glasgow, Scotland, UK, October. ACM.

Hal Daumé III. 2007. Frustratingly easy domain adaptation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 256–263, Prague, Czech Republic, June. Association for Computational Linguistics.

Pedro Domingos. 1999. MetaCost: a general method for making classifiers cost-sensitive. *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 155–164.

Benjamin Farber, Dayne Freitag, Nizar Habash, and Owen Rambow. 2008. Improving NER in Arabic using a morphological tagger. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odjik, Stelios Piperidis, and Daniel Tapias, editors, *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, pages 2509–2514, Marrakech, Morocco, May. European Language Resources Association (ELRA).

Radu Florian, Hany Hassan, Abraham Ittycheriah, Hongyan Jing, Nanda Kambhatla, Xiaoqiang Luo, Nicolas Nicolov, and Salim Roukos. 2004. A statistical model for multilingual entity detection and tracking. In Susan Dumais, Daniel Marcu, and Salim Roukos, editors, *Proceedings of the Human Language Technology Conference of the North*

American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004, page 18, Boston, Massachusetts, USA, May. Association for Computational Linguistics.

Radu Florian, John Pitrelli, Salim Roukos, and Imed Zitouni. 2010. Improving mention detection robustness to noisy input. In *Proceedings of EMNLP 2010*, pages 335–345, Cambridge, MA, October. Association for Computational Linguistics.

Dayne Freitag. 2004. Trained named entity recognition using distributional clusters. In Dekang Lin and Dekai Wu, editors, *Proceedings of EMNLP 2004*, pages 262–269, Barcelona, Spain, July. Association for Computational Linguistics.

Kevin Gimpel and Noah A. Smith. 2010a. Softmax-margin CRFs: Training log-linear models with loss functions. In *Proceedings of the Human Language Technologies Conference of the North American Chapter of the Association for Computational Linguistics*, pages 733–736, Los Angeles, California, USA, June.

Kevin Gimpel and Noah A. Smith. 2010b. Softmax-margin training for structured log-linear models. Technical Report CMU-LTI-10-008, Carnegie Mellon University. `http://www.lti.cs.cmu.edu/research/reports/2010/cmulti10008.pdf`.

Cyril Grouin, Sophie Rosset, Pierre Zweigenbaum, Karn Fort, Olivier Galibert, and Ludovic Quintard. 2011. Proposal for an extension of traditional named entities: from guidelines to evaluation, an overview. In *Proceedings of the 5th Linguistic Annotation Workshop*, pages 92–100, Portland, Oregon, USA, June. Association for Computational Linguistics.

Nizar Habash and Owen Rambow. 2005. Arabic tokenization, 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. 2010. *Introduction to Arabic Natural Language Processing*. Morgan and Claypool Publishers.

Ahmed Hassan, Haytham Fahmy, and Hany Hassan. 2007. Improving named entity translation by exploiting comparable and parallel corpora. In *Proceedings of the Conference on Recent Advances in Natural Language Processing (RANLP '07)*, Borovets, Bulgaria.

Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2006. OntoNotes: the 90% solution. In *Proceedings of the Human Language Technology Conference of the NAACL (HLT-NAACL)*, pages 57–60, New York City, USA, June. Association for Computational Linguistics.

Dirk Hovy, Chunliang Zhang, Eduard Hovy, and Anselmo Peas. 2011. Unsupervised discovery of domain-specific knowledge from text. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1466–1475, Portland, Oregon, USA, June. Association for Computational Linguistics.

Jing Jiang and ChengXiang Zhai. 2006. Exploiting domain structure for named entity recognition. In *Proceedings of the Human Language Technology Conference of the NAACL (HLT-NAACL)*, pages 74–81, New York City, USA, June. Association for Computational Linguistics.

Jun'ichi Kazama and Kentaro Torisawa. 2007. Exploiting Wikipedia as external knowledge for named entity recognition. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 698–707, Prague, Czech Republic, June. Association for Computational Linguistics.

Chloe Kiddon and Yuriy Brun. 2011. That's what she said: double entendre identification. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 89–94, Portland, Oregon, USA, June. Association for Computational Linguistics.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In Dekang Lin and Dekai Wu, editors, *Proceedings of EMNLP 2004*, pages 388–395, Barcelona, Spain, July. Association for Computational Linguistics.

LDC. 2005. ACE (Automatic Content Extraction) Arabic annotation guidelines for entities, version 5.3.3. Linguistic Data Consortium, Philadelphia.

Percy Liang, Hal Daumé III, and Dan Klein. 2008. Structure compilation: trading structure for features. In *Proceedings of the 25th International Conference on Machine Learning (ICML)*, pages 592–599, Helsinki, Finland.

Chris Manning. 2006. Doing named entity recognition? Don't optimize for F_1. `http://nlpers.blogspot.com/2006/08/doing-named-entity-recognition-dont.html`.

David McClosky, Eugene Charniak, and Mark Johnson. 2006. Effective self-training for parsing. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pages 152–159, New York City, USA, June. Association for Computational Linguistics.

Rada Mihalcea. 2004. Co-training and self-training for word sense disambiguation. In *HLT-NAACL 2004 Workshop: Eighth Conference on Computational Natural Language Learning (CoNLL-2004)*, Boston, Massachusetts, USA.

Einat Minkov, Richard Wang, Anthony Tomasic, and William Cohen. 2006. NER systems that suit user's preferences: adjusting the recall-precision trade-off for entity extraction. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, pages 93–96, New York City, USA, June. Association for Computational Linguistics.

Luke Nezda, Andrew Hickl, John Lehmann, and Sarmad Fayyaz. 2006. What in the world is a *Shahab*? Wide coverage named entity recognition for Arabic. In *Proccedings of LREC*, pages 41–46.

Joel Nothman, Tara Murphy, and James R. Curran. 2009. Analysing Wikipedia and gold-standard corpora for NER training. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2009)*, pages 612–620, Athens, Greece, March. Association for Computational Linguistics.

PediaPress. 2010. mwlib. `http://code.pediapress.com/wiki/wiki/mwlib`.

Slav Petrov, Pi-Chuan Chang, Michael Ringgaard, and Hiyan Alshawi. 2010. Uptraining for accurate deterministic question parsing. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 705–713, Cambridge, MA, October. Association for Computational Linguistics.

Lev Ratinov and Dan Roth. 2009. Design challenges and misconceptions in named entity recognition. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL-2009)*, pages 147–155, Boulder, Colorado, June. Association for Computational Linguistics.

Nathan D. Ratliff, J. Andrew Bagnell, and Martin A. Zinkevich. 2006. Subgradient methods for maximum margin structured learning. In *ICML Workshop on Learning in Structured Output Spaces*, Pittsburgh, Pennsylvania, USA.

Ryan Roth, Owen Rambow, Nizar Habash, Mona Diab, and Cynthia Rudin. 2008. Arabic morphological tagging, diacritization, and lemmatization using lexeme models and feature ranking. In *Proceedings of ACL-08: HLT*, pages 117–120, Columbus, Ohio, June. Association for Computational Linguistics.

Satoshi Sekine, Kiyoshi Sudo, and Chikashi Nobata. 2002. Extended named entity hierarchy. In *Proceedings of LREC*.

Burr Settles. 2004. Biomedical named entity recognition using conditional random fields and rich feature sets. In Nigel Collier, Patrick Ruch, and Adeline Nazarenko, editors, *COLING 2004 International Joint workshop on Natural Language Processing in Biomedicine and its Applications (NLPBA/BioNLP) 2004*, pages 107–110, Geneva, Switzerland, August. COLING.

Khaled Shaalan and Hafsa Raza. 2008. Arabic named entity recognition from diverse text types. In *Advances in Natural Language Processing*, pages 440–451. Springer.

Mihai Surdeanu, David McClosky, Mason R. Smith, Andrey Gusev, and Christopher D. Manning. 2011. Customizing an information extraction system to a new domain. In *Proceedings of the ACL 2011 Workshop on Relational Models of Semantics*, Portland, Oregon, USA, June. Association for Computational Linguistics.

Ben Taskar, Carlos Guestrin, and Daphne Koller. 2004. Max-margin Markov networks. In Sebastian Thrun, Lawrence Saul, and Bernhard Schölkopf, editors, *Advances in Neural Information Processing Systems 16*. MIT Press.

Antonio Toral, Elisa Noguera, Fernando Llopis, and Rafael Muñoz. 2005. Improving question answering using named entity recognition. *Natural Language Processing and Information Systems*, 3513/2005:181–191.

Ioannis Tsochantaridis, Thorsten Joachims, Thomas Hofmann, and Yasemin Altun. 2005. Large margin methods for structured and interdependent output variables. *Journal of Machine Learning Research*, 6:1453–1484, September.

Christopher Walker, Stephanie Strassel, Julie Medero, and Kazuaki Maeda. 2006. ACE 2005 multilingual training corpus. LDC2006T06, Linguistic Data Consortium, Philadelphia.

Ralph Weischedel and Ada Brunstein. 2005. BBN pronoun coreference and entity type corpus. LDC2005T33, Linguistic Data Consortium, Philadelphia.

Dan Wu, Wee Sun Lee, Nan Ye, and Hai Leong Chieu. 2009. Domain adaptive bootstrapping for named entity recognition. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 1523–1532, Singapore, August. Association for Computational Linguistics.

Tianfang Yao, Wei Ding, and Gregor Erbach. 2003. CHINERS: a Chinese named entity recognition system for the sports domain. In *Proceedings of the Second SIGHAN Workshop on Chinese Language Processing*, pages 55–62, Sapporo, Japan, July. Association for Computational Linguistics.