

Data Selection in Semi-supervised Learning for Name Tagging

Abstract

We present two semi-supervised learning techniques to improve a state-of-the-art multi-lingual name tagger. They improved F-measure 1.4% for English and 1.5% for Chinese. We also conclude that simply relying upon large corpora is not in itself sufficient: we must pay attention to unlabeled data selection too. We describe effective measures to automatically select documents and sentences.

1 Introduction

When applying machine learning approaches to natural language processing tasks, it is time-consuming and expensive to hand-label the large amounts of training data necessary for good performance. Unlabeled data can be collected in much larger quantities. Therefore, a natural question is whether we can use unlabeled data to build a more accurate learner, given the same amount of labeled data. This problem is often referred to as semi-supervised learning. It significantly reduces the effort needed to develop a training set. It has shown promise in improving the performance of many tasks such as name tagging (Miller et al., 2004), semantic class extraction (Lin et al., 2003), chunking (Ando and Zhang, 2005), coreference resolution (Bean and Riloff, 2004) and text classification (Blum and Mitchell, 1998).

However, it is not clear, when semi-supervised learning is applied to improve a learner, how the system should effectively select unlabeled data,

and how the size and relevance of data impact the performance.

In this paper we apply two semi-supervised learning algorithms to improve a state-of-the-art name tagger. We run the baseline name tagger on a large unlabeled corpus (bootstrapping) and the test set (self-training), and automatically generate high-confidence machine-labeled sentences as additional ‘training data’. We then iteratively re-train the model on the increased ‘training data’ until convergence.

We first investigated whether we can improve the system by simply using a lot of unlabeled data. By dramatically increasing the size of the corpus with unlabeled data, we did get a significant improvement compared to the baseline system. But we found that adding new data does not always help, and furthermore, even a manually review of some of the machine-tagged data yielded little further gain.

Then we tried to select relevant documents from the unlabeled data in advance, and got clear further improvements. We also obtained significant improvement by self-training (bootstrapping on the test data) without any additional unlabeled data.

Therefore, in contrast to the claim in (Banko and Brill, 2001), we concluded that, for some applications, effective use of large unlabeled corpora demands good data selection measures. We propose and quantify some effective measures to select documents and sentences in this paper.

The rest of this paper is structured as follows. Section 2 briefly describes the efforts made by previous researchers to use semi-supervised learning and well as the work of (Banko and Brill, 2001). Section 3 presents our baseline name tagger. Section 4 describes the motivations while Sec-

tion 5 presents the details of two semi-supervised learning methods. Section 6 presents and discusses the experimental results on both English and Chinese. Section 7 presents our conclusions and directions for future work.

2 Prior Work

This work presented here extends a substantial body of previous work (Blum and Mitchell, 1998; Riloff and Jones, 1999; Ando and Zhang, 2005) that all focus on reducing annotation requirements. For the specific task of named entity annotation, some researchers have emphasized the creation of taggers from minimal seed sets (Strzalkowski 1996; Collins and Singer 1999; Lin et al., 2003) while another line of inquiry (which we are pursuing) has sought to improve on high-performance baseline taggers (Miller et al., 2004).

Banko and Brill (2001) suggested that the development of very large training corpora may be most effective for progress in empirical natural language processing. Their experiments show a logarithmic trend in performance as corpus size increases without performance reaching an upper bound. Recent work has replicated their work on thesaurus extraction (Curran and Moens, 2002) and is-a relation extraction (Ravichandran et al., 2004), showing that collecting data over a very large corpus significantly improves system performance. However, (Curran, 2002) and (Curran and Osborne, 2002) claimed that the choice of statistical model is more important than relying upon large corpora.

3 Baseline Multi-lingual Name Tagger

Our baseline name tagger is based on an HMM that generally follows the Nymble model (Bikel et al, 1997). Within each of the name class states, a statistical bigram model is employed, with the usual one-word-per-state emission. The various probabilities involve word co-occurrence, word features, and class probabilities. Then it uses the best-first search to generate NBest hypotheses.

Since these probabilities are estimated based on observations seen in a corpus, “back-off models” are used to reflect the strength of support for a given statistic, as for the Nymble system.

The HMM tagger also computes the margin – the difference between the log probabilities of the

top two hypotheses. This is used as a rough measure of confidence in our name tagging.¹

In processing Chinese, to take advantage of name structures, we do name structure parsing using an extended HMM which includes a larger number of states (14). This new HMM can handle name prefixes and suffixes, and transliterated foreign names separately. It operates on the output of a word segmenter from Tsinghua University. We also augmented the HMM model with a set of post-processing rules to correct some omissions and systematic errors.

The name tagger identifies three name types: Person (PER), Organization (ORG) and GPE (geopolitical entities).

4 Semi-Supervised Learning for Name Tagging

The performance of name taggers has been limited in part by the amount of labeled training data available. How can an unlabeled corpus help to address this problem? Based on its original training (on the labeled corpus), there will be some tags (in the unlabeled corpus) that the tagger will be very sure about. For example, there will be contexts that were always followed by a person name (e.g., "Capt.") in the training corpus. If we find a new token T in this context in the unlabeled corpus, we can be quite certain it is a person name. If the tagger can learn this fact about T, it can successfully tag T when it appears in the test corpus without any indicative context. In the same way, if a previously-unseen context appears consistently in the unlabeled corpus before known person names, the tagger should learn that this is a predictive context.

We have adopted a simple learning approach: we take the unlabeled text about which the tagger has greatest confidence in its decisions, tag it, add it to the training set, and retrain the tagger. This process is performed repeatedly to bootstrap ourselves to higher performance. This approach can be used with any supervised-learning tagger that can produce some measure of confidence in its decisions.

¹ This metric was used by (Ji and Grishman, 2004) in the context of rescoring of name hypotheses; Scheffer et al. (2001) used a similar metric for active learning of name tags.

5 Two Semi-Supervised Learning Methods for Name Tagging

We have applied this bootstrapping approach to two sources of data: first, to a large corpus of unlabeled data and second, to the test set. To distinguish the two, we shall label the first "bootstrapping" and the second "self-training".

We begin (Sections 5.1 and 5.2) by describing the basic algorithms used for these two processes. We expected that these basic methods would provide a substantial performance boost, but our experiments showed that, for best gain, the additional training data should be related to the target problem, namely, our test set. We present measures to select documents (Section 5.3) and sentences (Section 5.4), and show (in Section 6) the effectiveness of these measures.

5.1 Bootstrapping

We divided the large unlabeled corpus into segments in order to: 1). create segments of manageable size; 2) separately evaluate the contribution of each segment (using a labeled development test set) and reject those which do not help; and 3) apply the latest updated best model to each subsequent segment. The procedure can be formalized as follows.

1. Split the unlabeled data into n subsets and mark them C_1, C_2, \dots, C_n . Call the updated HMM name tagger NameM (initially the baseline tagger), and a development test set DS.
2. For $i=1$ to n
 - (1) Run NameM on C_i ;
 - (2) For each tagged sentence S in C_i , If S is tagged with high confidence, then keep S ; otherwise remove S .
 - (3) Relabel the current name tagger (NameM) as OldNameM, add C_i to the training data, and re-train the name tagger, producing an updated model NameM.
 - (4) Run NameM on DS; if the performance gets worse, reset NameM = OldNameM

5.2 Self-training

An analogous approach can be used to tag the test set. The basic intuition is that the sentences in which the learner has low confidence may get sup-

port from those sentences previously labeled with high confidence.

Initially, we build the baseline name tagger from the labeled examples, then gradually add the most confidently tagged test sentences into the training corpus, and reuse them for next iteration until all sentences are labeled high confidently. The procedure can be formalized as follows.

1. Call the updated HMM name tagger NameM, and the test set TestS;
2. While (there are sentences tagged with high confidence)
 - (1) Run NameM on TestS;
 - (2) For each tagged sentence S in TestS, if S is tagged with high confidence, add S to the training data
 - (3) Re-train the name tagger, producing an updated model NameM.

Self-training takes more advantage of linguistic information for the task of name tagging. It can be considered a cache model variant, but with two more benefits: 1). It uses accurate confidence measures as weights for each name candidate, and relies on more accurate names to re-adjust the prediction of the remaining names, while in a cache model, all name candidates are equally weighted for voting (independent of the learner's confidence). 2). It adds high-confidence sentences and re-trains the learner so that all probabilistic parameters can be updated automatically, and the learner won't be biased by those high-confidence but incorrect tags.

5.3 Unlabeled Document Selection

To further investigate the benefits of using very large corpora in bootstrapping, and also inspired by the gain from self-training, we reconsidered the assumptions of our approach. The bootstrapping method implicitly assumes that the unlabeled data is reliable (not noisy) and useful, namely:

1. The unlabeled data supports the acquisition of new names and contexts, to reduce the sparse data problem;
2. The unlabeled data won't make the old estimates worse by adding too many names whose tags are incorrect, or at least are incorrect in

the context of the labeled training data and the test data.

If the unlabeled data is noisy or unrelated to the test data, it can hurt rather than improve the learner’s performance on the test set. So it is necessary to coarsely measure the relevance of the unlabeled data to our target test set, specifically for name tagging purposes. We define a relevance measure between the test set T and an unlabeled document d as follows.

(1) ‘Query set’ construction

We model the information expected from the unlabeled data by a ‘bag of words’ technique. We construct a query term set from the test set, and then use it to check whether each document in the unlabeled data is useful or not.

a). We prefer not to use all the words in the test set as key words, since we are only concerned about the distribution of name candidates. (Adding off-topic documents may in fact introduce noise into the model.) For example, if T talks about the presidential election in France while d talks about the presidential election in the US, they may share many common words such as ‘election’, ‘voting’, ‘poll’, and ‘camp’, but we would expect more gain from other unlabeled documents talking about the French election, since they may share many name candidates.

b). On the other hand it is insufficient to only take the name candidates in the first hypothesis for each sentence (since we are particularly concerned with tokens which *might* be names but are not so labeled in the top hypothesis). So our solution is to take all the name candidates in the top N best hypotheses for each sentence to construct a query set Q .

(2) Cross-entropy Measure

Using the query set Q , we computed the cross entropy $H(T, d)$ between the distribution of the name candidates in T and d . If $H(T, d)$ is smaller than a threshold then we consider d a useful unlabeled document².

$$H(T, d) = - \sum_{x \in Q} T(x) \cdot \log d(x)$$

² We also tried a single match method, using the query set to find all the relevant documents that include any names belonging to Q , and got approximately the same result as cross-entropy.

In addition to this relevance selection, we used one other simple filter: we removed a document if it includes fewer than five names, because it is unlikely to be news.

5.4 Sentence Selection

We don’t want to add all the generated sentences to the training corpus because incorrectly tagged or irrelevant sentences can lead to degradation in model performance. The value of larger corpora is partly dependent on how much new information is extracted from each sentence of the unlabeled data compared to the training corpus that we already have.

The following confidence measures were applied to assist the semi-supervised learning algorithm in selecting useful sentences for re-training the model.

1. Margin to find reliable sentences

For each sentence, we compute the HMM hypothesis margin (the difference in log probabilities) between the first hypothesis and the second hypothesis. We select the sentences with margins larger than a threshold to be added to the training data.

Unfortunately, the margin often comes down to whether a specific word has previously been observed in training; if the system has seen the word, it is certain, if not, it is uncertain. Therefore the sentences with high margins are a mix of interesting and uninteresting samples. We need to apply additional measures to remove the uninteresting ones. On the other hand, we may have confidence in a tagging due to evidence external to the HMM, so we explored measures beyond the HMM margin in order to recover additional sentences.

2. Name coreference to find more reliable sentences

For each document, we use simple coreference resolution between names such as string matching and name abbreviation resolution.

Assume S is one sentence in the document, and there are k names tagged in S : $\{N_1, N_2, \dots, N_k\}$, which are coreferred to by $\{CorefNum_1, CorefNum_2, \dots, CorefNum_k\}$ other mentions separately. Then we count the following average name coreference number *AveCoref* as a confidence measure:

$$AveCoref = (\sum_{i=1}^k CorefNum_i) / k$$

(Ji and Grishman, 2004) have shown that doing cross-document upon test set can provide a more reliable confidence measure about coreference numbers. In our experiments for self-training, we first use cross-entropy to cluster the related documents, and then apply cross-document name coreference for each cluster. Then we compute *AveCoref* as above, treating the entire cluster as a single document.

3. Name count and sentence length to remove uninteresting sentences

In bootstrapping on unlabeled data, by applying margin measure we often get some sentences which are too short or don't include any names. Although they are tagged with high confidence, they may make the model worse if added into the training data (for example, by artificially increasing the probability of non-names). In our experiments we don't use a sentence if it includes fewer than six words, or doesn't include any names.

5.5 Formal Framework

We formalize the above two semi-supervised learning methods in figure 1.

6 Evaluation Results and Discussions

6.1 Data

We evaluated our system on two languages: English and Chinese. Table 1 shows the data used in our experiments.

We present below the overall performance for both languages, and also some diagnostic experiment results.

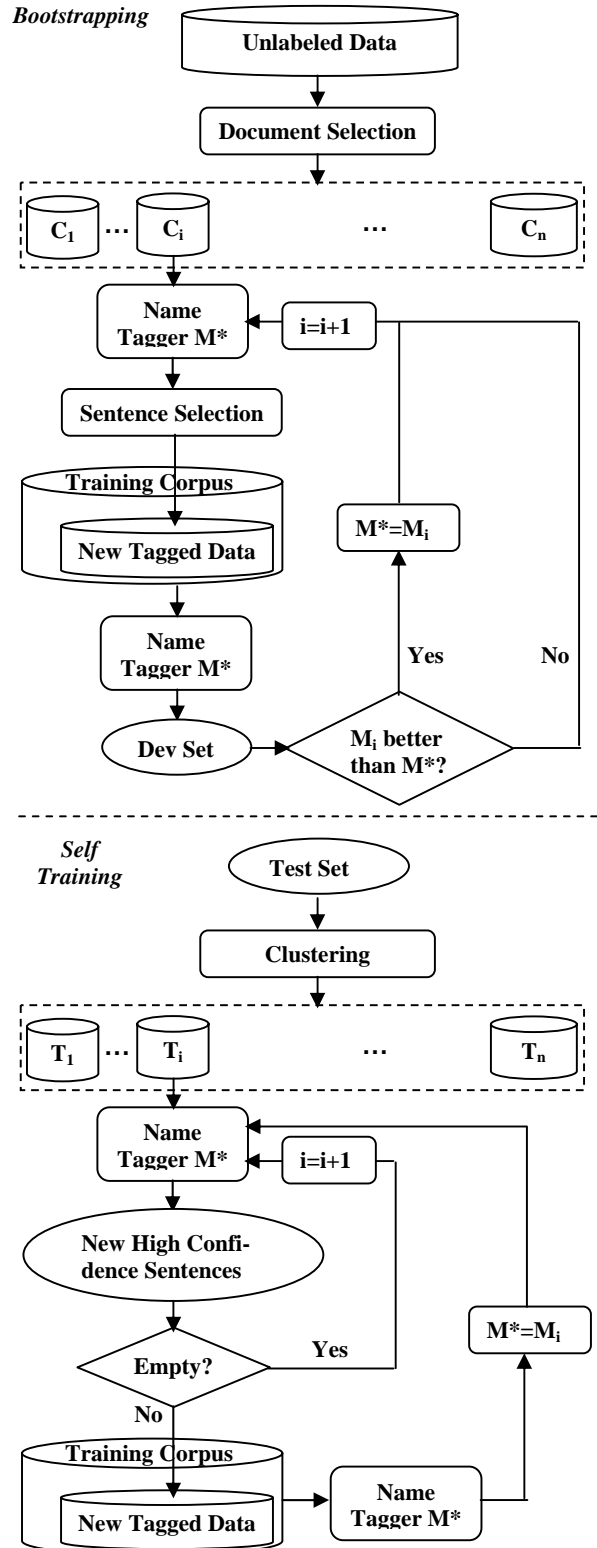


Figure 1. Semi-supervised Learning Frameworks for Name Tagging

Data		English	Chinese
Baseline Training data		ACE02,03,04 + BBN data 989,003 words	Beijing Corpus +ACE03,04,05 1,460,648 words
Unlabeled Data	Total	196,494 docs in Mar-Jun of 2003 (69M words)	14,402 docs in Nov&Dec of 2000 (9M words)
	Selected Docs	28,361 docs (557,123 sentences)	6,975 docs (83,275 sentences)
	Selected Sentences	127,247 sentences (2,745,176 words)	28,962 sentences (589,250 words)
Dev Set		20 ACE04 texts in Oct of 2000	90 ACE05 texts in Oct of 2000
Test Set		20 ACE04 texts in Oct of 2000	90 ACE05 texts in Oct of 2000

Table 1. Data Description

6.2 Overall Performance

Table 2 and Table 3 present the overall performance by applying the two semi-supervised learning methods, separately and in combination, to our baseline name tagger.

Learner	P	R	F
Baseline	88.8	89.3	89.0
Bootstrapping with data selection	89.2	90.1	89.6
Self-training	89.6	90.4	90.0
Bootstrapping with data selection + Self-training	90.2	90.6	90.4

Table 2. English Name Tagger

Learner	P	R	F
Baseline	88.2	87.6	87.9
Bootstrapping with data selection	88.9	89.0	88.9
Self-training	89.5	88.3	88.9
Bootstrapping with data selection + Self-training	89.6	89.3	89.4

Table 3. Chinese Name Tagger

We can see that the bootstrapping method obtained more gain for F-measure (F) for Chinese than English; while self-training performed more robustly on English. For English, the overall system achieves a 12.5% relative reduction on the spurious and incorrect tags, and 12.1% reduction in

the missing rate. For Chinese, it achieves a 11.9% relative reduction on the spurious and incorrect tags, and 13.7% reduction in the missing rate.

6.3 Analysis of Bootstrapping

1. Impact of Data Size

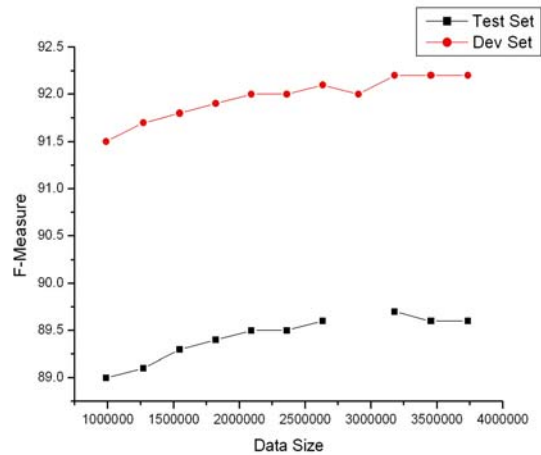


Figure 2. Impact of Data Size (English)

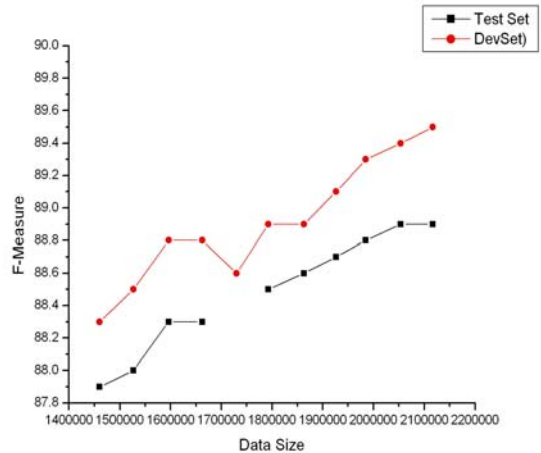


Figure 3. Impact of Data Size (Chinese)

Figure 2 and 3 show the results as each segment of the unlabeled data is added to the training corpus. We can see some flattening of the gain at the end, particularly for the larger English corpus, and that some segments (gaps in the test set graph) do not help to boost the performance.

2. Impact of Data Selection

In order to investigate the contribution of document selection in bootstrapping, we performed the following diagnostic experiments for Chinese.

	Learner	P	R	F
(1)	Baseline	88.2	87.6	87.9
(2)	Document Selection+ Sentence Selection + Bootstrapping	88.9	89.0	88.9
(3)	Bootstrapping	88.4	88.5	88.4
(4)	Bootstrapping + Manual Review	88.6	88.8	88.7
(5)	Single-Iteration Un- supervised Learning	88.2	88.4	88.3
(6)	Document Selection + Bootstrapping	88.8	88.2	88.5

Table 4. Impact of Data Selection (Chinese)

Comparing row 2 (with selection) with row 3 (without selection), we find that removing selection, even though it multiplies the size of the corpus, results in lower performance. The F-measure in row 3 is 0.5% lower than row 2. Adding off-topic material to the training data makes the performance worse.

We conducted a rapid manual review of the sentences generated by the row 3 method (and, we believe, corrected most errors), but little further gain (shown in row 4) is obtained. Notably, the performance is 0.2% lower than row 2 (with document selection and no manual review) and lower even than self-training, which does not use any additional data. This leads us to conclude that simply relying upon large corpora is not in itself sufficient. Effective use of large corpora demands good confidence measures for document selection.

The results in row 5 show that learning by incremental bootstrapping (using the segmented corpus) is more effective than treating the corpus as a whole and adding all sentences to the training set in a single iteration. Row 6 (without sentence selection) shows that good sentence selection methods are also important to improve the system.

6.4 Analysis of Self-training

1. Confidence Measures in Sentence Selection

We have tried different measures to extract high-confidence sentences. The results are presented in Figure 4.

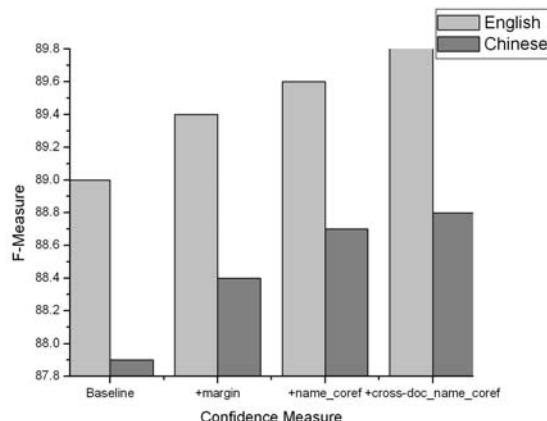


Figure 4. Impact of Confidence Measures

The graph shows that English benefits more from the clustering (cross-doc) method; Chinese benefits more from adding name coreference, mainly because there are more coreference links between name abbreviations and full names. And we also can see that the margin is the most crucial measure for both languages.

2. Margin Threshold Setting

We experimented with several different margin thresholds and found that gradually lowering the threshold so that 5% to 20% of the sentences (with the largest margin) were added to the training corpus at each iteration proved relatively effective. This yielded the following gradually improving performance for one English cluster including 7 documents and 190 sentences.

No. of iterations	No. of sentences added	No. of tags changed	F
0	0	0	91.4
1	37	28	91.9
2	69	22	92.1
3	107	21	92.4
4	128	11	92.6
5	146	9	92.7
6	163	8	92.8
7	178	6	92.8
8	190	0	92.8

Table 5. Incremental Improvement from Self-training (English)

7 Conclusions and Future Work

This paper demonstrates the effectiveness of two straightforward semi-supervised learning methods for improving a state-of-art name tagger, and investigates the importance of data selection for this application.

Banko and Brill (2001) suggested that the development of very large training corpora may be central to progress in empirical natural language processing. When using large amounts of unlabeled data, as expected, we did get improvement by using unsupervised bootstrapping. But our results did not entirely support the claim that exploring a very large corpus can by itself eliminate the sparseness problem. In contrast, we observed that good measures to select relevant unlabeled documents and useful labeled sentences are crucial.

In the future we plan to try topic identification techniques to select relevant unlabeled documents, and use the downstream information extraction components such as coreference resolution and relation detection to measure the confidence of the tagging for sentences. We are also interested in applying clustering as a pre-processing step for bootstrapping.

References

- Rie Kubota Ando and Tong Zhang. 2005. A High-Performance Semi-Supervised Learning Methods for Text Chunking. *Proc. ACL2005*. pp. 1-8. Ann Arbor, USA
- Michele Banko and Eric Brill. 2001. Scaling to very very large corpora for natural language disambiguation.. *Proc. ACL2001*. pp. 26-33. Toulouse, France
- David Bean and Ellen Riloff. 2004. Unsupervised Learning of Contextual Role Knowledge for Coreference Resolution. *Proc. HLT-NAACL2004*. pp. 297-304. Boston, USA
- Daniel M. Bikel, Scott Miller, Richard Schwartz, and Ralph Weischedel. 1997. Nymble: a high-performance Learning Name-finder. *Proc. Fifth Conf. on Applied Natural Language Processing, Washington, D.C.* pp. 194-201.
- Avrim Blum and Tom Mitchell. 1998. Combining Labeled and Unlabeled Data with Co-training. *Proc. of the Workshop on Computational Learning Theory*. Morgan Kaufmann Publishers
- Michael Collins and Yoram Singer. 1999. Unsupervised Models for Named Entity Classification. *Proc. of EMNLP/VLC-99*
- James R. Curran and Marc Moens. 2002. Scaling context space. *Proc. ACL 2002*, Philadelphia, USA
- James R. Curran. 2002. Ensemble Methods for Automatic Thesaurus Extraction. *Proc. EMNLP 2002*, Philadelphia, USA
- James R. Curran and Miles Osborne. 2002. A very very large corpus doesn't always yield reliable estimates. *Proc. ACL 2002 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics*, Philadelphia, USA
- Heng Ji and Ralph Grishman. 2004. Applying Coreference to Improve Name Recognition. *Proc. ACL 2004 Workshop on Reference Resolution and Its Applications*, Barcelona, Spain
- Winston Lin, Roman Yangarber and Ralph Grishman. 2003. Bootstrapping Learning of Semantic Classes from Positive and Negative Examples. *Proc. ICML-2003 Workshop on The Continuum from Labeled to Unlabeled Data*. Washington, D.C.
- Scott Miller, Jethran Guinness and Alex Zamanian. 2004. Name Tagging with Word Clusters and Discriminative Training. *Proc. HLT-NAACL2004*. pp. 337-342. Boston, USA
- Deepak Ravichandran, Patrick Pantel, and Eduard Hovy. 2004. The Terascale Challenge. *Proc. KDD Workshop on Mining for and from the Semantic Web (MSW-04)*. pp. 1-11. Seattle, WA, USA
- Ellen Riloff and Rosie Jones. 1993. Learning Dictionaries for Information Extraction by Multi-Level Bootstrapping. *Proc. AAAI/IAAI*
- Tobias Scheffer, Christian Decomain, and Stefan Wrobel. 2001. Active Hidden Markov Models for Information Extraction. *Proc. Int'l Symposium on Intelligent Data Analysis (IDA-2001)*.
- T. Strzalkowski and J. Wang. 1996. A Self-Learning Universal Concept Spotter. *Proc. COLING*.