Data Selection in Semi-supervised Learning for Name Tagging

Heng Ji **Ralph Grishman**
Department of Computer Science
New York University
New York, NY, 10003, USA
hengji@cs.nyu.edu grishman@cs.nyu.edu

Abstract

We present two semi-supervised learning techniques to improve a state-of-the-art multi-lingual name tagger. For English and Chinese, the overall system obtains 1.7% - 2.1% improvement in F-measure, representing a 13.5% - 17.4% relative reduction in the spurious, missing, and incorrect tags. 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'.

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 off-topic unlabeled data sometimes makes the performance worse. 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 as well as the work of (Banko and Brill, 2001). Section 3 presents our baseline name tagger. Section 4 describes the motivation for our approach while Section 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)

Proceedings of the Workshop on Information Extraction Beyond The Document, pages 48–55,
Sydney, July 2006. ©2006 Association for Computational Linguistics

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 and Wang, 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 Motivation

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 reliable measure of confidence in its decisions.

4 Baseline Multi-lingual Name Tagger

Our baseline name tagger is based on an HMM that generally follows the Nymble model (Bikel et al, 1997). Then it uses best-first search to generate NBest hypotheses, and 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. 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 Geopolitical (GPE) entities (locations which are also political units, such as countries, counties, and cities).

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 based on news sources and dates 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

[1] We have also used this metric in the context of rescoring of name hypotheses (Ji and Grishman, 2005); Scheffer et al. (2001) used a similar metric for active learning of name tags.

segment. The procedure can be formalized as follows.

1. Select a related set *RelatedC* from a large corpus of unlabeled data with respect to the test set *TestT*, using the document selection method described in section 5.3.

2. Split *RelatedC* into n subsets and mark them C_1, C_2…C_n. Call the updated HMM name tagger *NameM* (initially the baseline tagger), and a development test set *DevT*.

3. 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 retrain the name tagger, producing an updated model *NameM*;

   (4) Run *NameM* on *DevT*; if the performance gets worse, don't use C_i and 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 support 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 the next iteration, until all sentences are labeled. The procedure can be formalized as follows.

1. Cluster the test set *TestT* into n clusters T_1, T_2, …,T_n, by collecting document pairs with low cross entropy (described in section 5.3.2) into the same cluster.

2. For i=1 to n
   (1) *NameM* = baseline HMM name tagger;

   (2) While (there are new sentences tagged with confidence higher than a threshold)
      a. Run *NameM* on T_i;
      b. Set an appropriate threshold for margin;
      c. For each tagged sentence *S* in T_i, if *S* is tagged with high confidence, add *S* to the training data;
      d. Retrain the name tagger *NameM* with augmented training data.

At each iteration, we lower the threshold so that about 5% of the sentences (with the largest margin) are added to the training corpus.[2] As an example, 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-Measure
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 1. Incremental Improvement from Self-training (English)

Self-training can be considered a cache model variant, operating across the entire test collection. But it uses confidence measures as weights for each name candidate, and relies on names tagged with high confidence 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).

5.3 Unlabeled Document Selection

To further investigate the benefits of using very large corpora in bootstrapping, and also inspired by the gain from the "essence" of self-training, which aims to gradually emphasize the predictions from *related* sentences within the test set, we reconsidered the assumptions of our approach. The bootstrapping method implicitly assumes that the unlabeled data is reliable (not noisy) and uniformly useful, namely:

[2] To be precise, we repeatedly reduce the threshold by 0.1 until an additional 5% or more of the sentences are included; however, if more than an additional 20% of the sentences are captured because many sentences have the same margin, we add back 0.1 to the threshold.

- The unlabeled data supports the acquisition of new names and contexts, to provide new evidence to be incorporated in HMM and reduce the sparse data problem;
- 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. We define an IR (information retrieval) - style relevance measure between the test set *TestT* and an unlabeled document *d* as follows.

5.3.1 '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 corpus *TestT* to check whether each unlabeled document *d* is useful or not.

- We prefer not to use all the words in *TestT* 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 one document in *TestT* 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.

- On the other hand it is insufficient to only take the name candidates in the top one 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*.

5.3.2 Cross-entropy Measure

Using *Q*, we compute the cross entropy *H(TestT, d)* between *TestT* and *d* by:

$$H(TestT,d) = -\sum_{x \in Q} prob(x \mid TestT) \times \log_2 prob(x \mid d)$$

where *x* is a name candidate in *Q*, and *prob(x|TestT)* is the probability (frequency) of *x* appearing in *TestT* while *prob(x|d)* is the probability of *x* in *d*. If *H(T, d)* is smaller than a threshold then we consider *d* a useful unlabeled document[3].

5.4 Sentence Selection

We don't want to add all the tagged sentences in a relevant document 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.

5.4.1 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[4] 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.

---

[3] 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.

[4] In bootstrapping, this margin threshold is selected by testing on the development set, to achieve more than 93% F-Measure.

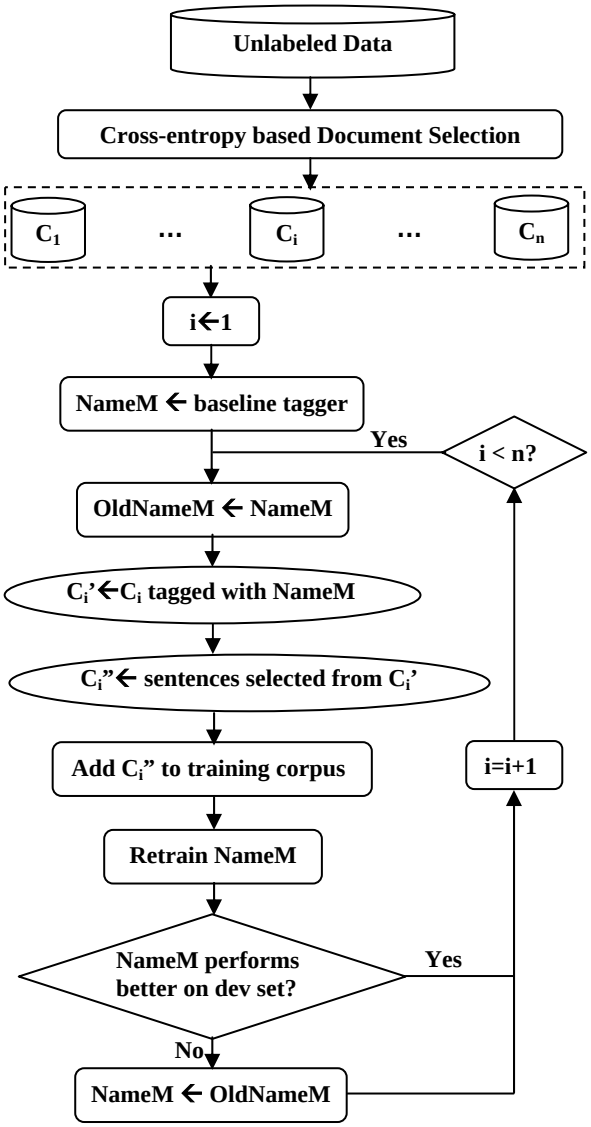

Figure 1. Bootstrapping for Name Tagging

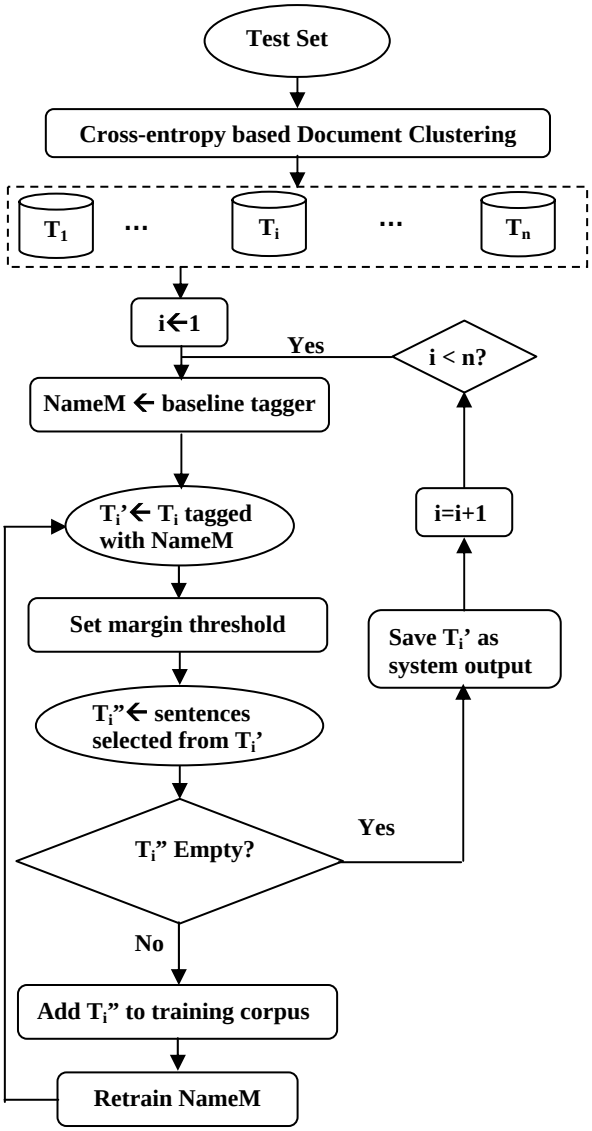

Figure 2. Self-Training for Name Tagging

Data		English	Chinese
Baseline Training data		ACE02,03,04 989,003 words	Beijing Corpus +ACE03,04,05 1,460,648 words
Unlabeled Data	Total	196,494 docs in Mar-Jun of 2003 (69M words) from ACE05 unlabeled data	41061 docs in Nov,Dec of 2000, and Jan of 2001 (25M words) from ACE05 and TDT4 transcripts
	Selected Docs	62584 docs (1,314,148 Sentences)	14,537 docs (222,359 sentences)
	Selected Sentences	290,973 sentences (6,049,378 words)	55,385 sentences (1,128,505 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 and 80 ACE05 texts in Mar-May of 2003 (3093 names, 1205 PERs, 1021GPEs, 867 ORGs)	90 ACE05 texts in Oct of 2000 (3093 names, 1013 PERs, 695 GPEs, 769 ORGs)

Table 2. Data Description

5.4.2 Name coreference to find more reliable sentences

Names introduced in an article are likely to be referred to again, so a name coreferred to by more other names is more likely to have been correctly tagged. In this paper, we use simple coreference resolution between names such as substring matching and name abbreviation resolution.

In the bootstrapping method we apply single-document coreference for each individual unlabeled text. In self-training, in order to further benefit from global contexts, we consider each cluster of relevant texts as one single big document, and then apply cross-document coreference.

Assume S is one sentence in the document, and there are k names tagged in S: $\{N_1, N_2 \ldots\ldots N_k\}$, which are coreferred to by $\{CorefNum_1, CorefNum_2, \ldots CorefNum_k\}$ other names separately. Then we use the following average name coreference count *AveCoref* as a confidence measure for tagging S:[5]

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

5.4.3 Name count and sentence length to remove uninteresting sentences

In bootstrapping on unlabeled data, the margin criterion often selects 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 Data Flow

We depict the above two semi-supervised learning methods in Figure 1 and Figure 2.

6 Evaluation Results and Discussions

6.1 Data

We evaluated our system on two languages: English and Chinese. Table 2 shows the data used in our experiments.

[5] For the experiments reported here, sentences were selected if *AveCoref* > 3.1 (or 3.1×number of documents for cross-document coreference) *or* the sentence margin exceeded the margin threshold.

We present in section 6.2 – 6.4 the overall performance of precision (P), recall (R) and F-measure (F) for both languages, and also some diagnostic experiment results. For significance testing (using the sign test), we split the test set into 5 folders, 20 texts in each folder of English, and 18 texts in each folder of Chinese.

6.2 Overall Performance

Table 3 and Table 4 present the overall performance[6] by applying the two semi-supervised learning methods, separately and in combination, to our baseline name tagger.

Learner	P	R	F
Baseline	**87.3**	**87.6**	**87.4**
Bootstrapping with data selection	**88.2**	**88.6**	**88.4**
Self-training	**88.1**	**88.4**	**88.2**
Bootstrapping with data selection + Self-training	**89.0**	**89.2**	**89.1**

Table 3. English Name Tagger

Learner	P	R	F
Baseline	**88.2**	**87.6**	**87.9**
Bootstrapping with data selection	**89.8**	**89.5**	**89.6**
Self-training	**89.5**	**88.3**	**88.9**
Bootstrapping with data selection + Self-training	**90.2**	**89.7**	**90.0**

Table 4. Chinese Name Tagger

For English, the overall system achieves a 13.4% relative reduction on the spurious and incorrect tags, and 12.9% reduction in the missing rate. For Chinese, it achieves a 16.9% relative reduction on the spurious and incorrect tags, and 16.9% reduction in the missing rate.[7] For each of the five folders, we found that both bootstrapping and self-training produced an improvement in F score for each folder, and the combination of two methods is always better than each method alone. This allows us to reject the hypothesis that these

[6] Only names which exactly match the key in both extent and type are counted as correct; unlike MUC scoring, no partial credit is given.

[7] The performance achieved should be considered in light of human performance on this task. The ACE keys used for the evaluations were obtained by dual annotation and adjudication. A single annotator, evaluated against the key, scored F=93.6% to 94.1% for English and 92.5% to 92.7% for Chinese. A second key, created independently by dual annotation and adjudication for a small amount of the English data, scored F=96.5% against the original key.

improvements were random at a 95% confidence level.

6.3 Analysis of Bootstrapping

6.3.1 Impact of Data Size

Figure 3 and 4 below show the results as each segment of the unlabeled data is added to the training corpus.

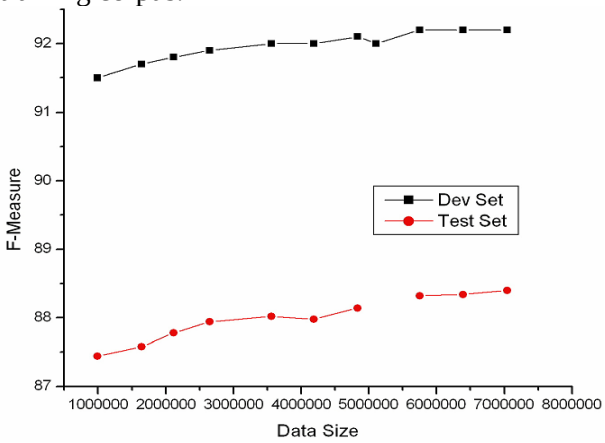

Figure 3. Impact of Data Size (English)

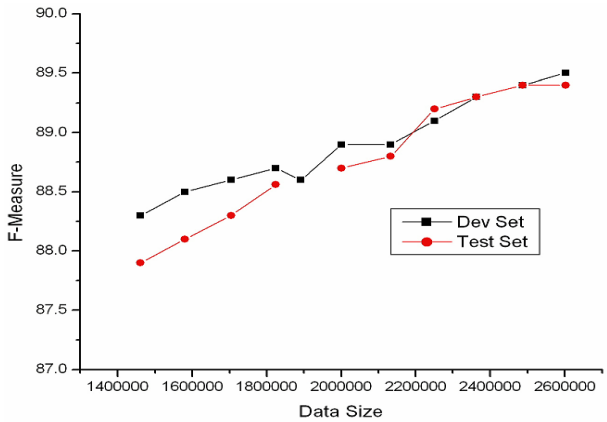

Figure 4. Impact of Data Size (Chinese)

We can see some flattening of the gain at the end, particularly for the larger English corpus, and that some segments do not help to boost the performance (reflected as dips in the Dev Set curve and gaps in the Test Set curve).

6.3.2 Impact of Data Selection

In order to investigate the contribution of document selection in bootstrapping, we performed diagnostic experiments for Chinese, whose results are shown in Table 5. All the bootstrapping tests (rows 2 - 4) use margin for sentence selection; row 4 augments this with the selection methods described in sections 5.4.2 and 5.4.3.

Learner		P	R	F
(1)	Baseline	88.2	87.6	87.9
(2)	(1) + Bootstrapping	88.9	88.7	88.8
(3)	(2) + Document Selection	89.3	88.9	89.1
(4)	(3) + Sentence Selection	89.8	89.5	89.6

Table 5. Impact of Data Selection (Chinese)

Comparing row 2 with row 3, we find that not using document selection, even though it multiplies the size of the corpus, results in 0.3% lower performance (0.3-0.4% loss for each folder). 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 to remove off-topic material. By adding sentence selection (results in row 4) the system obtained 0.5% further improvement in F-Measure (0.4-0.7% for each folder). All improvements are statistically significant at the 95% confidence level.

6.4 Analysis of Self-training

We have applied and evaluated different measures to extract high-confidence sentences in self-training. The contributions of these confidence measures to F-Measure are presented in Table 6.

Confidence Measure	English	Chinese
Baseline	87.4	87.9
Margin	87.8	88.3
Margin + single-doc name coreference	88.0	88.7
Margin + cross-doc name coreference	88.2	88.9

Table 6. Impact of Confidence Measures

It shows that 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 an important measure for both languages. All differences are statistically significant at the 95% confidence level except for the gain using cross-document information for the Chinese name tagging.

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. However, exploiting a very large corpus did not by itself produce the greatest performance gain. Rather, we observed that good measures to select relevant unlabeled documents and useful labeled sentences are important.

The work described here complements the active learning research described by (Scheffer et al., 2001). They presented an effective active learning approach that selects "difficult" (small margin) sentences to label by hand and then add to the training set. Our approach selects "easy" sentences – those with large margins – to add automatically to the training set. Combining these methods can magnify the gains possible with active learning.

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 preprocessing step for bootstrapping.

Acknowledgment

This material is based upon work supported by the Defense Advanced Research Projects Agency under Contract No. HR0011-06-C-0023, and the National Science Foundation under Grant IIS-00325657. Any opinions, findings and conclusions expressed in this material are those of the authors and do not necessarily reflect the views of the U. S. Government.

References

Rie 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.* pp.194-201. Washington D.C., USA

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. 2005. Improving Name Tagging by Reference Resolution and Relation Detection. *Proc. ACL2005*. pp. 411-418. Ann Arbor, USA.

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. 1999. 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).*

Tomek Strzalkowski and Jin Wang. 1996. A Self-Learning Universal Concept Spotter. *Proc. COLING.*