Generating Chinese Named Entity Data from a Parallel Corpus

Ruiji Fu, Bing Qin, Ting Liu*
Research Center for Social Computing and Information Retrieval
MOE-Microsoft Key Laboratory of Natural Language Processing and Speech
School of Computer Science and Technology
Harbin Institute of Technology, Harbin, China
{rjfu,bqin,tliu}@ir.hit.edu.cn

Abstract

Annotating Named Entity Recognition (NER) training corpora is a costly process but necessary for supervised NER systems. This paper presents an approach to generate large-scale Chinese NER training data from an English-Chinese discourse level aligned parallel corpus. Difficulty of NER is different among languages due to their unique features. For example, the performance of English NER systems is usually higher than the Chinese ones on average. In our method, we first employ a high performance NER system on one side of a bilingual corpus. And then, we project the NE labels to the other side according to the word level alignment. At last, we select high-quality labeled sentences using different strategies and generate an NER training corpus. In our experiments, we generate a Chinese NER corpus with 167,100 sentences through an English-Chinese parallel corpus. The system trained on the automatically generated corpus attains a comparable result with the one trained on the manually-annotated corpus. Further experiments show that the NER performance is significantly improved on two different evaluation sets by using the generated training data as an additional corpus to the manually-labeled data.

1 Introduction

Named Entity Recognition (NER) is the task of identifying and classifying the names of persons, locations, organizations and other named entities in text, which plays an important role in many Natural Language Processing (NLP) applications such as information extraction, information retrieval, machine translation, and so on.

Supervised machine learning systems have proved successful for NER (Zhou and Su, 2002; Chieu and Ng, 2002; Takeuchi and Collier, 2002; Settles, 2004). They usually need manually-annotated high performance textual corpora. These corpora are considered as gold standards for training statistical models. However, corpora manually-annotating is so costly and time-consuming that the existing corpora are limited in both scale and scope for Chinese NER.

More seriously, the domain overfitting problem even worsens the corpora-shortage problem. Supervised NER approaches can often achieve high accuracy when a large annotated training set similar to the test data is available (Zhou and Su, 2002; Florian et al., 2003; Klein et al., 2003; Finkel et al., 2005). Unfortunately, if the test data has some difference from the training data, these approaches tend to not perform well. For instance, Ciaramita and Altun (2005) reported that the F1-score of a named entity recognizer trained on CoNLL 2003 Reuters corpus dropped from 90.8% (when tested on a similar Reuters set) to 64.3% (when tested on a Wall Street Journal set). A similar phenomenon of performance degradation in Chinese NER will be presented later in this paper (see Table 3).

Therefore, we try to solve the problems for Chinese mentioned above by automatically constructing large scale and scope training corpora.

Chinese NER is more difficult than English NER because of the lack of capitalization and the uncertainty in word segmentation. Our motivation is to collect large-scale training data and improve Chinese NER with the help of an exist-

* Correspondence author: tliu@ir.hit.edu.cn

Proceedings of the 5th International Joint Conference on Natural Language Processing, pages 264–272,
Chiang Mai, Thailand, November 8 – 13, 2011.

ing high performance English NER system and a bilingual corpus.

In this paper, we employ a high performance NER system on the English side of a bilingual corpus. And then, the NE labels are projected to the Chinese side according to the word level alignment. At last, we select high-quality labeled sentences using different strategies and generate an NER training corpus.

In our experiments, statistical models are trained on the generated corpora, and compared with the model trained on a manually annotated corpus. The results show that our corpus is comparable to the manually-labeled corpus. Furthermore, the model trained on the combined corpus (generated and manually-labeled corpora) obtains an F1-score of 67.89% on 863-Evaluation corpus and 73.20% on OntoNotes corpus, which significantly outperforms the one trained on the manually-labeled corpus.

The contributions of this paper are as follows.

First, we present a method to generate large-scale Chinese NER training data from a bilingual corpus automatically. Our method trades off manual effort to annotate named entities in documents for effort to identify pairs of parallel documents, which is easier than NE manual annotation. For example, large scale of parallel documents can be extracted from the web automatically (Resnik and Smith, 2003; Zhang et al., 2006).

Second, we propose some strategies to select high-quality training data, which are very effective and important as the experiments show.

And third, we prove that our generated training data can be used as an additional corpus to improve the NER performance.

This paper is organized as follows. Section 2 discusses the related work. Section 3 describes our approach in detail. Section 4 presents and discusses the results of our experiments. Finally, we present our conclusions and future work in section 5.

2 Related Work

In this section, we introduce some previous work about NER training data generation.

The most closed related work to our approach is Yarowsky et al. (2001). They used word alignment on parallel corpora to induce several text analysis tools from English to other languages for which such resources are scarce. An NE tagger was transferred from English to French and achieved good classification accuracy. However, Chinese NER is more difficult than French and word alignment between Chinese and English is also more complex because of the tremendous difference between the two languages.

Huang and Vogel (2002) presented an integrated approach to extract an NE translation dictionary from an English-Chinese parallel corpus while improving the monolingual NE annotation quality for both languages. They started with low-quality NE tagging for both languages and improved the annotation result using alignment information. But they did not filter the annotated data and evaluate its impact for NER as training data.

Besides, some other resources have been used to generate NE tagged corpus.

An et al. (2003) and Whitelaw et al. (2008) used seed sets of entities and search engines to collect NER training data from the web. However, constructing of a high-quality seed list is also a time-consuming work.

Richman and Schone (2008) and Nothman et al. (2008) used similar methods to create NE training data. They transformed Wikipedia's links into named entity annotations by classifying the target articles into common entity types. But the article classification seeds also had to be hand-labeled in advance.

In the biomedical domain, Vlachos and Gasperin (2006) automatically created training material for the task of gene name recognition from the broader raw corpus using existing domain resources.

In our work, we generate a large scale Chinese NER training data from a bilingual corpus without any NE seed lists and filter it by using effective strategies. And we prove that it can improve the performance of Chinese NER as additional training data.

3 Our Approach

In this section we describe our approach of generating NER training data from a parallel corpus. The framework of our system consists of four components as shown in Figure 1.

- **Alignment:** Sentence alignment and word alignment is performed on a discourse-level aligned bilingual corpus.
- **English NER:** We identify NEs on the English side of the parallel corpus, making use of an existing high performance English NER system.
- **NE Candidates Generation:** Based on the result of the word alignment, we

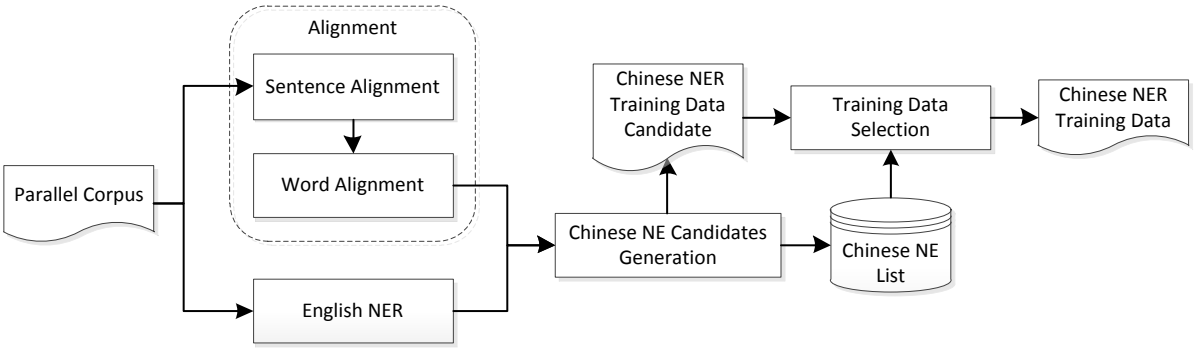

Figure 1. System Framework

project the English NE labels to the Chinese side and generate training data candidates. At the same time, we extract a Chinese NE list, which can be used as a dictionary resource.

- **Training Data Selection:** According to the different filtering strategies, we select high-quality labeled sentences from the candidates to form Chinese NER training data.

3.1 Alignment and Automatic English NER

First, we perform sentence level alignment by using Champollion toolkit[1].

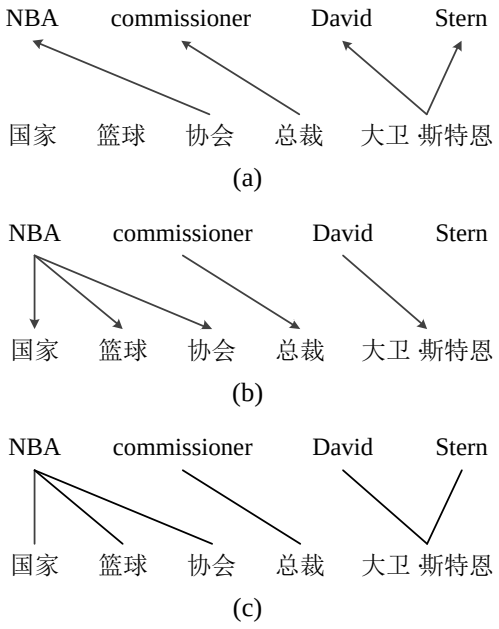

Figure 2. (a) Word Alignment from Chinese to English. (b) Word Alignment from English to Chinese. (c) The Merged Result of Both Directions. In (a), *国家* and *篮球* are aligned to NULL, the same to *Stern* in (b).

[1] http://champollion.sourceforge.net/

And then, GIZA++ toolkit[2] is used for word alignment. This toolkit can generate one-to-many word alignments in a certain direction (Chinese to English or English to Chinese). However, we need many-to-many alignments. Hence, we need GIZA++ to run on the bilingual corpus in both directions and merge the results, as shown in Figure 2.

English NER is easier than Chinese because of the capitalization information and the needlessness of word segmentation. So the performance of English NER systems is usually higher than the Chinese ones on average. Hence a widely used open-source NER system, Stanford Named Entity Recognizer[3] is employed to label NEs on the English side of the parallel corpus. The system is based on linear chain Conditional Random Field (CRF) (J.Lafferty et al., 2001) sequence models and can recognize three kinds of named entities (PERSON, LOCATION and ORGANIZATION).

To evaluate the robustness of Stanford NER system, we manually labeled 1000 English sentences from our bilingual corpus as a test set where the system achieves an F1-score of 89.32%. It is close to the result of 87.94%[4] on CoNLL 2003 NER test set.

3.2 Chinese NE Candidates Generation

After the English NER, we map the English NE labels to the Chinese side to discover Chinese NEs candidates, according to the result of word alignment.

We consider all related alignment pairs of every word within an English NE. For example, in Figure 3, the index of the organization name *NBA* is 1 and the related word alignment pairs

[2] http://www-i6.informatik.rwth-aachen.de/Colleagues/och/software/GIZA++.html
[3] http://nlp.stanford.edu/software/CRF-NER.shtml
[4] http://nlp.stanford.edu/projects/project-ner.shtml

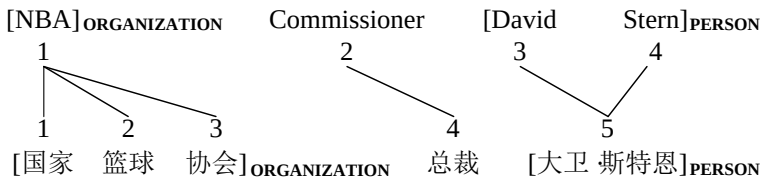

Figure 3. An Example of Chinese NE Candidates Generation

include 1-1, 1-2 and 1-3. Therefore, we can find the boundaries (from 1 to 3) of the corresponding Chinese translation *国家 篮球 协会*. There are also some English words connecting with NULL at Chinese side. We ignore these word alignment pairs.

According to the alignment, we project the NE labels from English to Chinese and generate the named entity candidates on the Chinese side.

3.3 Training Data Selection

However, the generated NER training data candidates are noisy because of the errors in English NER or word alignment. In this section, we present the strategies of selecting training data.

3.3.1 Filtering Based on Rules

As the common definition, a named entity is a continuous string, whether it is in English or in Chinese. So we assume that every named entity alignment pair is a closed alignment pair of two continuous strings, as shown in Figure 4 (a).

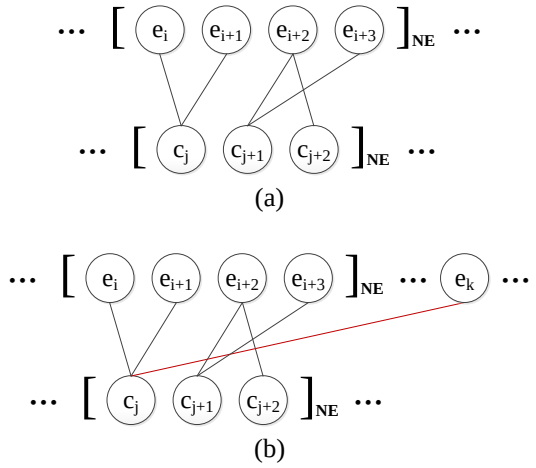

Figure 4. (a) An eligible case; (b) An ineligible case. In (b), the word alignment pair e_k - c_j is against the rule, while k > i+3 or k < i.

Based on this assumption, we make two alternative rules to filter the training data candidates. One is a soft filtering rule to retain training instances as many as possible. Another is a hard filtering rule to guarantee the quality of the generated corpus. These two rules are shown as follows:

- **Rule 1 (the soft rule):** Label a Chinese NE candidate as a non-NE, if a word within it has an alignment pair with an English word out of the corresponding English NE, such as Figure 4 (b).
- **Rule 2 (the hard rule):** Discard the whole sentence where there is a case satisfying Rule 1.

Rule 1 prefers to keep training instances as many as possible. But it may make some NEs be labeled as non-NE mistakingly on the Chinese side for incorrect word alignments, which are the noises in the generated training data. Rule 2 prefers to guarantee the quality of the generated data but may make useful training instances be discarded and the data scale shrinking.

Based on the rules, we can filter lots of ill conditioned named entity candidates, such as overlapped entities, nested entities and so on.

3.3.2 Filtering Based on Scores

Although many ill conditioned candidates are filtered out by the rules, the remaining data is still noisy because of the incorrect labeling of the English NER and the incorrect NE alignment. In fact, the accuracy of NE alignment is only affected by the boundary alignment of English and Chinese NEs. In other words, we do not care about how to align within or without the NEs. Hence, we score Chinese named entity candidates by formula 1.

$$score(N_c) = \varphi(N_e) \prod_{w \in B(N_c)} \left(\frac{1}{|A(w)|} \sum_{\langle e,w \rangle \in A(w)} p(\langle e,w \rangle) \right) \quad (1)$$

Here, $\varphi(N_e)$ denotes the confidence of the English named entity N_e, which is derived from Stanford NER system. $B(N_c)$ denotes the boundaries of the Chinese named entity N_c, which are actually the left-most and the right-most word within N_c. e denotes an English word, and w denotes a Chinese word. $A(w)$ denotes all related alignment pairs of word w in current Chinese

named entity N_c. $p(\langle e,w\rangle)$ denotes the probability of alignment $\langle e,w\rangle$, which is obtained from GIZA++.

As mentioned in section 3.1, Stanford NER is based on CRF. The inference of CRF is that given an observable sequence $\vec{x}$, we want to find the most likely set of labels $\vec{y}$ for $\vec{x}$. The probability of $\vec{y}$ given $\vec{x}$ is calculated as follows (J.Lafferty et al., 2001):

$$p(\vec{y}|\vec{x}) = \frac{1}{Z(\vec{x})}\prod_{j=1}^{n}\psi_j(\vec{x},\vec{y}) \quad (2)$$

$$Z(\vec{x}) = \sum_{\vec{y}'}\prod_{j=1}^{n}\psi_j(\vec{x},\vec{y}') \quad (3)$$

$$\psi_j(\vec{x},\vec{y}) = exp\left(\sum_{i=1}^{m}\lambda_i f_i(\mathrm{y}_{j-1},y_j,\vec{x},j)\right) \quad (4)$$

In formulae 2, 3 and 4, j denotes the index of the jth word in sequence $\vec{x}$. n denotes the length of $\vec{x}$. m denotes the number of the features.

Now the substring $x_k x_{k+1}\cdots x_{k+l}$ in $\vec{x}$ is labeled as an NE N_e. The label sequence of N_e is $y_k^* y_{k+1}^*\cdots y_{k+l}^*$ which is denoted as $\vec{y}_{N_e}^*$. We compute the marginal probability $\varphi(N_e)$ as follows:

$$\varphi(N_e) = \frac{Z(N_e,\vec{x})}{Z(\vec{x})} \quad (5)$$

$$Z(N_e,\vec{x}) = \sum_{\vec{y}':\, y'_{k\ldots k+l}=\vec{y}_{N_e}^*}\prod_{j=1}^{n}\psi_j(\vec{x},\vec{y}') \quad (6)$$

The factor $\varphi(N_e)$ of every English NE is used to measure the confidence of NER. We apply the forward-backward algorithm to compute them.

For $p(\langle e,w\rangle)$, we use the probabilities of alignment pairs which are computed by GIZA++. GIZA++ outputs the probability $p(t|s)$ of translating source word *s* as target word *t*. There are two kinds of probabilities of alignment in two directions. Since our alignment is bidirectional, we merge the probabilities in two directions to come up with formula 7.

$$p(\langle e,w\rangle) = max\{p(e|w),p(w|e)\} \quad (7)$$

Particularly we set $p(t|s)$ zero while the translation pair "s → t" does not exist in the translation table given by GIZA++.

We set experiential thresholds for every category to filter the Chinese NE candidates.

3.3.3 Recalling by a Chinese NE List

During the time of filtering the NE candidates, we can also extract the high-quality candidates as an NE list. We calculate the frequencies and the average scores of the candidates. We set thresholds of frequency and average score for every kind of NE candidate, and select the candidates with the highest frequency and score to compose a list. Table 1 shows some samples of the list.

An NE may be found correctly in some sentences where word alignment is easy, while the same one may be missed in others. Hence we use the extracted NE list to recall the missed NEs in the result of the former two steps.

NEs	Label	Average Score	Freq.
北京 (Beijing)	LOC	0.637	4615
克林顿 (Clinton)	PER	0.853	969
联合国 (UN)	ORG	0.471	436
台湾 海峡 (Taiwan Strait)	LOC	0.244	82

Table 1. Samples of the Chinese NE List

4 Experiments

We carried out experiments to investigate the quality and practical applicability of our NER training corpora generated from the bilingual corpus.

4.1 Data Set

We selected the LDC2003E14 multilanguage corpus and several bilingual parallel corpora[5] as the source corpus to generate NER training data. LDC2003E14 was derived from news of Foreign Broadcast Information Service (FBIS). We used the English-Chinese parallel news composed of 11,645 document pairs. The other bilingual parallel corpora contain 11,750 sentence pairs in all.

A manually annotated Chinese NER gold-standard data from People's Daily corpus was prepared as the contrasting data. The corpus was annotated with 7 tags: person, location, organization, date, time, number and miscellany. For the evaluation, the last 4 tags were removed. The corpus, composed of 47,426 sentences, was di-

[5] Six Chinese-English sentence-aligned corpora were used as extra data, including LDC2002T01, LDC2003E04, E07, E08, T17 and LDC2004T07.

vided into two parts: 37,426 for training and 10,000 for evaluation.

We also use other two corpora for the evaluation. One is the Chinese NER evaluation corpus from the National High Technology Development 863 Program of China in 2004. The other is OntoNotes Release 2.0 corpus. The tags except person, location and organization were removed in the 863-Evaluation corpus. The OntoNotes corpus was annotated with 18 fine-grained tags: 11 for named entities and 7 for numerical and time terms. We reduced the tags *NORP* and *LOCATION* into location, *FACILITY*, *GPE* and *ORGANIZATION* into organization, and *PERSON* into person. After this preprocessing, 14,547 NEs remained in the 863-Evaluation corpus and 13,658 NEs remained in the OntoNotes corpus. See Table 2 for a summary of the corpora used.

Corpus	# of sentences	
	TRAIN	TEST
People's Daily	37,426	10,000
863-Evaluation	---	3,923
OntoNotes	---	6,904

Table 2. Corpora Used for Evaluation

In addition, we manually labeled 1,000 sentences randomly extracted from FBIS corpus for the direct evaluation about the quality of our generated corpus.

4.2 The Baseline System

We trained a Maximum Entropy Markov Model (MEMM) on People's Daily training set as our baseline. State-of-the-art features (Wu et al., 2005) are used, which contain word features, POS features, position features, and labeled NE tag features. The model was evaluated on the People's Daily test set, 863-Evaluation corpus and OntoNotes corpus. The result is shown in Table 3.

	P	R	F1
People's Daily	90.77%	88.90%	89.82%
863-Evaluation	74.55%	59.13%	65.87%
OntoNotes	78.32%	64.28%	70.56%

Table 3. Evaluation Result of the Baseline System

From the evaluation result, we can see that the F1-score drops from 89.82% on People's Daily corpus to 65.87% on 863-Evaluation corpus and to 70.56% on OntoNotes corpus. It's similar to the report of Ciaramita and Altun (2005). The reason for this problem is that the model is overfitted to the training data and fails to fit the test data with different distribution. To ease the problem, we attempt to improve the coverage of the model by generating large scale and scope training corpora.

4.3 The Quality of the Generated Data

We evaluated the training data generated by using different strategies on the 1000 manually annotated sentences.

	Size	P	R	F1
Rule1 only	1,000	76.16%	48.56%	59.30%
Rule1+Score	1,000	76.36%	48.49%	59.32%
Rule1+List	1,000	73.99%	67.71%	70.71%
Rule1+Score + List	1,000	74.28%	67.65%	70.81%
Rule2 only	661	78.61%	74.76%	76.64%
Rule2+Score	661	**78.77%**	74.76%	76.72%
Rule2+List	661	78.06%	86.44%	82.04%
Rule2+Score +List	661	78.19%	**86.44%**	**82.11%**

Table 4. The Quality of Generated Corpora

Comparing the upper and lower parts of Table 4, we get a larger corpus based on Rule 1, but the recall rate is low. Rule 2 requires removing whole sentences with ineligible cases, so that we can get a higher quality but smaller corpus. The result is reasonable. If an ineligible case as shown in Figure 4 (b) occurs in a pair of English and Chinese sentences, it is possible that a named entity is labeled in the English sentence, but is not mapped to the correct Chinese string due to the errors of word alignment. For example, if *National Basketball Association* is recognized as an organization name in an English sentence, it is very possible that the translated organization name *国家篮球协会* exists in the Chinese sentence. But if *National Basketball Association* is not aligned to *国家篮球协会*, the Chinese NE will be missed. We should remove the whole sentences from the corpus, or they will be noises in the training data. The results show that Rule 2 outperforms Rule 1. In the remainder of our experiment, we use Rule 2 instead of Rule 1.

The strategy filtering candidates by scores can help to improve the precision. But the improvement of F1-score is marginal, because some correct training instances may be filtered out, which makes the recall rate decrease.

Training data	863-Evaluation corpus			OntoNotes corpus		
	P	R	F1	P	R	F1
People's Daily (PD)	74.55%	59.13%	65.87%	78.32%	64.28%	70.56%
Rule2 only	72.94%	41.73%	53.09%	77.79%	46.90%	58.52%
Rule2+Score	73.66%	41.53%	53.11%	78.42%	46.72%	58.56%
Rule2+List	72.64%	58.08%	64.55%	76.84%	61.80%	68.50%
Rule2+Score+List	73.04%	58.12%	64.73%	76.90%	61.82%	68.54%
Rule2+Score+List+PD	**75.95%**	**61.38%**	**67.89%**	**80.35%**	**67.22%**	**73.20%**

Table 5. Test Results for the Generated Training Data

We extract a Chinese NE list containing 824 NEs with the highest frequency and scores from the corpus. Based on the NE list, we can recall many NEs missed by other strategies with only a little expense of precision. The recall rates are substantially improved from 48.56% to 67.71% based on Rule 1 and from 74.76% to 86.44% based on Rule 2.

Here, we chose the best-performing thresholds of the strategies in our experiments. The results show that our strategies are effective for improve the quality of the training data.

As mentioned in section 3.1, the F1-score of Stanford NER on our parallel corpus is 89.32%. The best F1-score of the generated training data in Table 4 is 82.11%. Thus, we roughly infer that about 8.07% (= 1 - 82.11% / 89.32%) correct NE information is lost in the process of the Chinese NER training data generation.

4.4 Comparison between the Manually-labeled Data and the Generated Data

We trained MEMMs on the generated corpora using the same features as the baseline. As shown in Table 5, we get a basic result by using Rule 2. On 863-Evaluation corpus for example, we get a marginal raise (0.72%) of precision but a drop (0.20%) of recall by using Rule 2 and Score strategy, and get a substantial raise (16.35%) of recall with a drop (0.30%) of the precision by using Rule 2 and List strategy. The situation is similar on OntoNotes corpus. The results are consistent with the quality of the training data shown in Table 4. And it is reasonable that better training data leads to higher NER performance. The model trained on our corpus generated by using all of the strategies gets a comparable result with the baseline system.

We also use the generated corpus as additional training data to the gold-standard data. The last row in Table 5 shows that this approach leads to an improvement of the NER performance. We also perform a paired significance test[6], which shows that the improvement is significant.

Our generated corpus contains 167,100 sentences, which are much more than sentences in the baseline corpus. Furthermore, it is generated without any manual annotation. The size could be limitless as long as there are plenty of parallel corpora available.

4.5 The Effect of the Generated Data Size

Figure 5 illustrates the effect of varying the size of the generated training data set. Increased training data tends to improve performance until the size reaches about 67k sentences for 863-Evaluation corpus and 33k for OntoNotes corpus. But after that, improvements are marginal.

We believe that there are two reasons causing this result. On the one hand, the noises increase when more training data is used. And the training data gets a balance between the noises and the correct training instances when the size reaches a certain point. On the other hand, a subset of the training data can represent the whole data, especially for the data from a simplex source. So we should collect data from a wider range of sources.

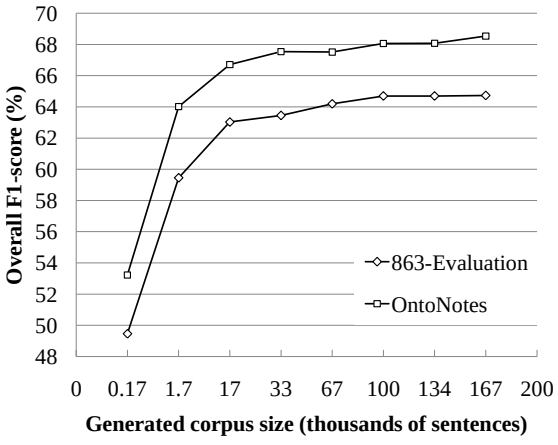

Figure 5. The Effect of Varying the Generated Corpus Size

[6] We used Zhang's significance tester (Zhang et al., 2004).

Training data	863-Evaluation corpus			OntoNotes corpus		
	PER	LOC	ORG	PER	LOC	ORG
People's Daily (PD)	59.91%	74.97%	35.61%	72.96%	76.89%	47.72%
Rule2+Score+List	62.68%	72.36%	28.81%	66.36%	76.43%	44.78%
Rule2+Score+List+PD	**66.15%**	**75.00%**	**36.22%**	**75.24%**	**78.31%**	**55.10%**

Table 6. Test Results for Each NE Category

4.6 The Performance of Each NE Category

To analyze overall error, our per-class F1-score is shown in Table 6. Training on the combined corpus could improve the performance of each NE category. The improvements are substantial in all categories except LOC on 863-Evaluation.

In general, the results of ORG entities are lower than the results of PER and LOC. The possible reason may be that ORG names are more complex than PER and LOC names. They usually consist of more words, which may result in more word alignment errors and then lead to more training instances filtered out. Fewer training instances might lead to a poorer performance. In addition, English ORG entity recognition is also more difficulty, which also results in more noises among the ORG name training instances.

5 Conclusion and Future Work

To solve the data-shortage and domain overfitting problems, we attempt to enlarge the Chinese NE training data automatically.

In this paper, we present a method of generating NER training data automatically from a bilingual parallel corpus. We employ an existing high-performance English NER system to recognized NEs at the English side, and then project the labels to the Chinese side according to the word alignment. To guarantee the quality of the training data, we propose effective filtering strategies. The results show that our training data is comparable with the manually-labeled data and can improve the performance of NER as an additional corpus.

Besides, the training data could be expanded easily as long as there are plenty of parallel corpora available. And identifying pairs of parallel documents is much easier than NE training data annotation. Generating training data from parallel corpora thus provides an alternative way of collecting data required for Chinese NER. Our method can be easily adapted to other languages.

In the future, we will try to improve the entity alignment and propose other better filtering strategies. Moreover, we will try to make use of more parallel corpora from a wider range of sources, because more parallel corpora may improve the accuracy of word alignment and widen the coverage of the generated NE corpus.

Acknowledgments

This work was supported by National Natural Science Foundation of China (NSFC) via grant 60975055 and 61073126. Special thanks to Wanxiang Che, Yanyan Zhao, Fikadu Gemechu, Yuhang Guo, Zhenghua Li, Meishan Zhang and the anonymous reviewers for insightful comments and suggestions.

References

Joohui An, Seungwoo Lee, and Gary Geunbae Lee. 2003. Automatic acquisition of named entity tagged corpus from world wide web. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 165-168.

Hai Leong Chieu and Hwee Tou Ng. 2002. Named Entity Recognition: A Maximum Entropy Approach Using Global Information. In *Proceedings of the 19th International Conference on Computational Linguistics (COLING)*, pages 190-196.

Massimiliano Ciaramita and Yasemin Altun. 2005. Named-entity recognition in novel domains with external lexical knowledge. In *Workshop on Advances in Structured Learning for Text and Speech Processing (NIPS-2005)*.

Jenny Finkel, Shipra Dingare, Christopher D. Manning, Malvina Nissim, Beatrice Alex, and Claire Grover. 2005. Exploring the boundaries: Gene and protein identification in biomedical text. *BMC Bioinformatics, 6(Suppl 1):S5*.

Radu Florian, Abe Ittycheriah, Hongyan Jing, and Tong Zhang. 2003. Named entity recognition through classifier combination. In *Proceedings of the 7th Conference on Natural Language Learning (CoNLL)*, pages 168-171.

Fei Huang and Stephan Vogel. 2002. Improved Named Entity Translation and Bilingual Named Entity Extraction. In *Proceedings of the 4th IEEE International Conference on Multimodal Interface*, pages 253-258. Pittsburgh, PA, October.

Dan Klein, Joseph Smarr, Huy Nguyen, and Christopher D. Manning. 2003. Named entity recognition with character-level models. In *Proceedings of the 7th Conference on Natural Language Learning (CoNLL)*, pages 188-191.

John Lafferty, Andrew McCallum and Fernando Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *International Conference on Machine Learning (ICML)*, pages 282-289.

Joel Nothman, James R. Curran, and Tara Murphy. 2008. Transforming Wikipedia into named entity training data. In *Proceedings of the Australian Language Technology Workshop*, pages 124-132, Hobart, Australia.

Philip Resnik and Noah A. Smith, 2003. The Web as a Parallel Corpus. *Computational Linguistics*, v.29 n.3, pages 349-380,

Alexander E. Richman and Patrick Schone. 2008. Mining wiki resources for multilingual named entity recognition. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1-9, Columbus, Ohio, USA.

Burr Settles. 2004. Biomedical Named Entity Recognition Using Conditional Random Fields and Rich Feature Sets. In *Proceedings of COLING 2004, the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications (NLPBA)*, Geneva, Switzerland.

Koichi Takeuchi and Nigel Collier. 2002. Use of Support Vector Machines in Extended Named Entity Recognition. In *Proceedings of the 6th Conference on Natural Language Learning (CoNLL)*, pages 119-125.

Andreas Vlachos and Caroline Gasperin. 2006. Bootstrapping and evaluating named entity recognition in the biomedical domain. In *Proceedings of the Workshop on Linking Natural Language Processing and Biology: Towards Deeper Biological Literature Analysis*, pages 138-145. New York City, New York, USA.

Casey Whitelaw , Alex Kehlenbeck , Nemanja Petrovic and Lyle Ungar. 2008. Web-scale named entity recognition. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, October 26-30, Napa Valley, California, USA.

Youzheng Wu, Jun Zhao, Bo Xu, Chinese Named Entity Recognition Model Based on Multiple Features. In *Proceedings of HLT/EMNLP 2005*, pages: 427-434, October 6-8, Vancouver, B.C., Canada.

David Yarowsky, Grace Ngai and Richard Wicentowski. 2001. Inducing multilingual text analysis tools via robust projection across aligned corpora. In *Human Language Technology Conference*, pages 109-116, San Diego, California, March.

Ying Zhang, Ke Wu, Jianfeng Gao, and Phil Vines. 2006. Automatic Acquisition of Chinese–English Parallel Corpus from the Web. In *Proceedings of ECIR-06, 28th European Conference on Information Retrieval*.

Ying Zhang, Stephan Vogel, and Alex Waibel. 2004. Interpreting BLEU/NIST scores: How much improvement do we need to have a better system? In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC)*, pages 2051-2054.

Guodong Zhou and Jian Su. 2002. Named entity recognition using an HMM-based chunk tagger. In *Proceedings of the 40th Annual Meeting of the Association of Comparative Linguistics (ACL)*, pages 473-480.