A Semi-Supervised Approach to Build Annotated Corpus for Chinese Named Entity Recognition

Xiaoshan FANG[#], Jianfeng GAO[*], Huanye SHENG[#]
[*]Microsoft Research Asia, jfgao@microsoft.com
[#]Shanghai Jiaotong University, China, {fang-xs, hysheng}@sjtu.edu.cn

Abstract

This paper presents a semi-supervised approach to reduce human effort in building an annotated Chinese corpus. One of the disadvantages of many statistical Chinese named entity recognition systems is that training data may be in short supply, and manually building annotated corpus is expensive. In the proposed approach, we construct an 80M hand-annotated corpus in three steps: (1) Automatically annotate training corpus; (2) Manually refine small subsets of the automatically annotated corpus; (3) Combine small subsets and whole corpus in a bootstrapping process. Our approach is tested on a state-of-the-art Chinese word segmentation system (Gao et al., 2003, 2004). Experiments show that only a small subset of hand-annotated corpus is sufficient to achieve a satisfying performance of the named entity component in this system.

1 Introduction

The success of applying statistical methods to natural language processing tasks depends to a large degree upon the quality and amount of available training data.

This paper presents our method of creating training data for the statistical Chinese word segmenter proposed in Gao *et al*. (2003). The segmenter is based on improved source-channel models, which are trained on a large amount of annotated training data. Whereas the hand-annotation is a very expensive task, creating the training data automatically remains an open research problem. Our approach falls somewhere between the two extremes of the spectrum. We try to minimize the human effort while keeping the quality of the annotation reasonably good for model estimation. The method to be presented has been discussed briefly in Gao *et al.* (2003). This paper presents an extended description with more details and experimental results.

The training data refer to a set of Chinese sentences where word boundaries and types have been annotated. Our basic solution is the bootstrapping approach described in Gao *et al*. (2002). It consists of three steps: (1) Initially, we use a greedy word segmenter to annotate the corpus, and obtain initial models based on the initial annotated corpus; (2) We re-annotate the corpus using the obtained models; (3) Re-train the models using the re-annotated corpus. Steps 2 and 3 are iterated until the performance of the system converges.

In this approach, the quality of the resulting models depends to a large degree upon the quality of the initial annotated corpus. Because there are many named entities that are not stored in a dictionary, traditional dictionary-based forward maximum matching (FMM) algorithm is not sufficient to create a good initial corpus. We thus manually annotate named entities on a small subset (call *seed set*) of the training data. Then, we obtain a model on the seed set (called *seed model*). We thus improve the initial model which is trained on the initial annotated training corpus by interpolating it with the seed model. Our experiments show that a relatively small seed set (e.g., 10 million characters, which takes approximately three weeks for 4 persons to annotate the NE tags) is enough to get a good improved model for initialization.

The remainder of this paper is organized as follows: Section 2 summarizes the related work. Section 3 deals with our approach to improve model estimation for Chinese word segmentation. The experiments are presented at Section 4. Finally we conclude in Section 5.

2 Related work

Traditional statistical approaches use a parametric model with maximum likelihood estimation (MLE), usually with smoothing methods to deal with data sparseness problems. These approaches have been introduced for the task of Chinese word segmentation. According to the training data used (word-segmented or not), the Chinese word segmentation can be achieved in a supervised or unsupervised manner.

As an example of unsupervised training, Ge *et al.* (1999) presents a simple zero-th order Markov

[1] This work was done while the author was visiting Microsoft Research Asia.

model of the words in Chinese text. They developed an efficient algorithm to train their model on an unsegmented corpus. Their basic assumption is that Chinese words are usually 1 to 4 characters long. They however did not take into account a large amount of named entities (e.g. Chinese organization name, transliterate name and some person names) most of which are longer than 4 characters (e.g., 微软亚洲研究院 Microsoft Research Asia, 加利福尼亚 California, 陈欧阳晓彤 a woman's name which puts her husband's surname ahead).

An and Wong used Hidden Markov Models (HMM) for segmentation. Their system is solely trained on a corpus which has been manually annotated with word boundaries and Part-of-Speech tags. Wu (2003) also used the training data to tune the segmentation parameters of their MSR-NLP Chinese system. He used the annotated training data to deal with the morphologically derived words.

In this paper we present a semi-supervised training method where we use both an auto-segmented training corpus and a small hand-annotated subset of it. Comparing to unsupervised approaches, our approach leads to a better segmenter that can identify much more named entities which are not in the dictionary. Comparing to supervised approaches, our method requires much less human effort for data annotation.

The Chinese word segmenter used in this study is described in Gao *et al*. (2003). The segmenter provides a unified approach to word segmentation and named entity (NE) recognition. This unified approach is based on the improved source-channel models of Chinese sentence generation, with two components: a source model and a set of channel models. For each word class (e.g. a person name), there is a channel model (referred to as *class model* afterwards) that estimates the generative probability of a character string given the word type. The source model is used to estimate the generative probability of a word sequence, in which each word belongs to one word class (e.g. a word in a lexicon or a named entity). In another word, it indicates, given a context, how likely a word occurs. So the source model is also referred to as *context model* afterwards. This paper focuses the discussion on how to create annotated corpus for context model estimation.

3 A semi-supervised approach to improve context model estimation

In this study the context model is a trigram model which estimates the probability of a word class.

Ideally, given an annotated corpus, where each sentence is segmented into words which are tagged by their word types, the trigram word class probabilities can be calculated using MLE, together with a backoff schema (Katz, 1987) to deal with the sparse data problem. Unfortunately, building such annotated training corpora is very expensive.

Our basic solution is the bootstrapping approach described in Gao et al. (2002). It consists of three steps: (1) Initially, a greedy word segmenter (i.e. FMM) is used to annotate the corpus, and an initial context model is obtained based on the initial annotated corpus; (2) Re-annotate the corpus using the obtained models; (3) Re-train the context model using the re-annotated corpus. Steps 2 and 3 are iterated until the performance of the system converges.

In the above approach, the quality of the context model depends to a large degree upon the quality of the initial annotated corpus, which is however not satisfied due to the fact that many named entities cannot be identifying using the greedy word segmenter which is based on the dictionary. As a consequence, the above approach achieves a low accuracy in detecting Chinese named entities.

A straightforward solution to the above problem is to obtain large amount of high-quality annotated corpus for context model estimation. Unfortunately, manually creating such annotated corpus is very expensive. For example, Douglas (1999) pointed out that at least up to about 1.2 million words of training data are necessary to train an HMM name recognizer. To guarantee a high degree of accuracy (e.g. 90% F-measure), it requires about 800 hours, or 20 person*weeks of labor to annotate and check the amount of data. This is almost certainly more time than would be required by a skilled rule writer to write a rule-based name recognizer achieving the same level of performance, assuming all the necessary resources, such as lexicons and name lists, are already available.

Our training data contains approximately 80 million Chinese characters from various domains of text. We are facing three questions in annotating the training data. (1) How to generate a high quality hand-annotated corpus? (2) How to best use the valuable hand-annotated corpus so as to achieve a satisfying performance? (3) What is the optimal size of the hand-annotated corpus, considering the tradeoff between the cost of human labor and the performance of the resulting segmenter?

We leave the answers to the first and third questions to Section 4. In what follows, we describe our method of using small set of human-annotated corpus to boost the quality of the annotation of the entire corpus. It consists of 6 steps.

Step 1: Manually annotate named entities on a small subset (call seed set) of the training data.

Step 2: Obtain a context model on the seed set (called seed model).

Step 3: Re-annotate the training corpus using the seed model and then train an improved context model using the re-annotated corpus.

Step 4: Manually annotate another small subset of the training data. Repeat Steps (2) and (3) until the entire training data have been annotated.

Step 5: Repeat steps 1 to 4 using different seed sets (we used three seed sets in our experiments, as we shall describe in Section 4).

Step 6: Combine all context models obtained in step 5 via linear interpolation:

$$P(xyz) = \Sigma\ \lambda_i \times P_i(xyz) \tag{1}$$

Here $P_i(xyz)$ is the trigram probability of the i-th context model. λ_s is the interpolation weights which vary from 0 to 1.

4 Experiments

In this section, we first present our experiments on the generation and evaluation of hand-annotated corpus to answer the first two questions. Then, the answer to the third question is given in subsection 4.2.

4.1 The generation and evaluation of hand-annotated corpus

4.1.1 The generation of hand-annotated corpus

Four students, whose major is Chinese language, annotate the corpus according to a pre-defined MSRA's guideline of Chinese named entities. We find that we have to revise the guideline when they were annotating the corpus. For example, Chinese character string "申城博览会 (Shanghai Exposition)"can be tagged as either "[L 申]城博览会" or "[L 申城]博览会". Here "申" is the abbreviation of "上海(Shanghai)". "城" is the abbreviation of "城市(city)". L is the tag of location name. It is not clearly described in the guideline where the named entity's right boundary is.

We obtain in total three manually annotated subsets (i.e. seed sets) by the following process:

1. Annotate the training data using a greedy word segmenter. Highlight the NEs and their tags.
2. Randomly select 10 million characters from the annotated training data and then ask the students to manually refine these 10 million characters. The refinement includes correcting the wrong NE tags and adding missing NE tags.
3. Repeat the second step, and then combine the obtained new 10-million-character subset with the first one. Hence, a 20-million-character subset of the training data is obtained.
4. Repeat the second step, and then combine the obtained new 10-million-character subset with the 20-million-character subset. Hence, a 30-million-character subset of the training data is obtained.

A manually annotated test set was developed as well. The text corpus contains approximately a half million Chinese characters that have been proofread and balanced in terms of domain, styles, and times.

4.1.2 The evaluation of hand-annotated corpus

To evaluate the quality of our annotated corpus, we trained a context model using the method described in Section 3, with the first-obtained 10-million-character seed set. We then compare the performance of the resulting segmenter with those of other state-of-the-art segmenters and the FMM segmenter.

4.1.2.1 Evaluation metrics

We conduct evaluations in terms of precision (P) and recall (R).

$$P = \frac{number \cdot of \cdot correctly \cdot identified \cdot NEs}{number \cdot of \cdot identified \cdot NEs} \tag{2}$$

$$R = \frac{number \cdot of \cdot correctly \cdot identified \cdot NEs}{number \cdot of \cdot all \cdot NEs} \tag{3}$$

4.1.2.2 Segmenters in Comparison

1. The **MSWS** system is one of the best available products. It is released by Microsoft® (as a set of Windows APIs). **MSWS** first conducts the word-breaking using MM (augmented by heuristic rules for disambiguation), and then conducts factoid detection and NER using rules.
2. The **LCWS** system is one of the best research systems in mainland China. It is released by Beijing Language University. The system works similarly to **MSWS**, but has a larger dictionary containing more PNs and LNs.
3. The **PBWS** system is a rule-based Chinese parser which can also output the word segmentation results. It explores high-level linguistic knowledge, such as syntactic structure for Chinese word segmentation and NER.

4.1.2.3 Results

The performance of the resulting segmenter is compared with those of three state-of-the-art segmenters and FMM segmenter in Table 1. Here PN, LN and ON stand for person name, location name and organization name respectively. The first column lists the segmenters.

As can be seen from Table 1, the resulting segmenter (SSSC.10m) achieves comparable results with those of the other three state-of-the-art word segmenters. From Table 1 we also find that our semi-supervised approach makes a 2.4%-49% improvement over FMM.

Segmenter	PN		LN		ON	
	R %	P %	R %	P %	R %	P %
MSWS	74.4	90.7	44.2	93.5	46.9	64.2
LCWS	78.1	94.5	72.0	85.4	13.1	71.3
PBWS	78.7	78.0	73.6	76.7	21.6	81.7
FMM	65.7	84.4	82.7	76.0	56.6	38.6
SSSC.10m	73.6	86.6	80.7	89.5	84.3	56.8
Impr. (%)	12.0	2.6	2.4	17.7	49.0	47.1

Table 1: Results on different Chinese word segmenters

The results show a moderate amount of hand-annotated corpus leads our segmenter to a state-of-the-art performance.

4.2 The optimal size of the hand-annotated corpus

Regarding the third question: what is the optimal size of the hand-annotated subset, considering the tradeoff between the cost of human labor and the performance of the resulting segmenter? We obtain a series of results using 10-30-million-character subsets as seed sets, and then plot three graphs showing the relationship between the performances and their corresponding human efforts of constructing the seed set.

4.2.1 Baselines

We use two baselines.

One is the FMM method, which does not use annotated training data. It is used to evaluate the performances using 10-30-million-character subsets (see Figure 1).

The other is to use all the human effort of annotating the whole training data, which takes about 1920 person*hours human effort. It is used to calculate how much labor we would save by using the semi-supervised approach described in section 3.

4.2.2 Results

The relationship between the performances and their corresponding human efforts of constructing the seed sets is shown in Figure 1. The X-axes give the human efforts on building 10, 20, and 30-million-character subsets. They are 360, 720, and 1080 person*hours. The Y-axes show the recall and precision results on person name, location name and organization name, separately.

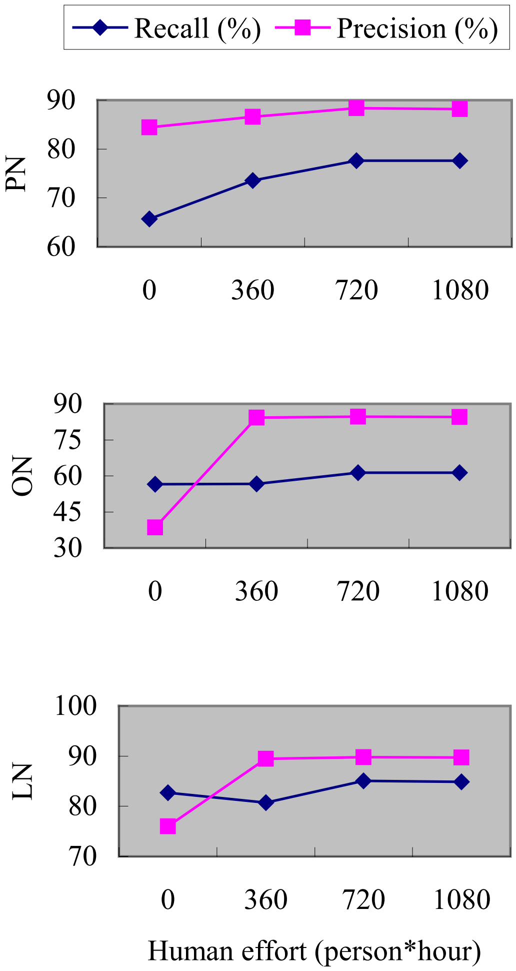

Figure 1: The relationship between the performances and their corresponding human efforts

We observe that both the recall and precision results first go upwards, and level off after the use of 720 person*hours, which is the corresponding human effort of constructing 20 million characters.

This means that 20 million characters is a saturation point, because more human effort does not lead to any improvement in performance, and less human effort leads to lower performance.

From the fact that manually annotating the whole training data costs 1920 person*hours, we indicate that by using our semi-supervised approach we save 62.5% human labor in corpus annotation.

5 Conclusion

This paper presents a semi-supervised method to save human effort in building annotated corpus. This method uses a small set of human-annotated corpus to boost the quality of the annotation of the entire corpus. We test this method on Gao's Chinese word segmentation system, which achieves a state-of-the-art performance on SIGHAN backoff data sets (Gao et al, 2004).

Several conclusions can be drawn from our experiments:

- The obtained corpus is of high quality.
- 20-million-characters is the optimal size of hand-annotated subset to boost the 80-million-character training data, considering the trade-off between the cost of human labor and the performance of the resulting segmenter.
- We save 62.5% human labor in corpus annotation.

References

An, Q. and Wong, W. S. 1996. *Automatic segmentation and tagging of Hanzi text using a hybrid algorithm.* In Proceedings of the 9th International Conference on Indus-trial & Engineering Applications of AI & Expert Systems.

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

Douglas Appelt. 1999. *Introduction to Information Extraction Technology.* A Tutorial Prepared for IJCAI-99.

Gao, Jianfeng, Joshua Goodman, Mingjing Li and Kai-Fu Lee. 2002. *Toward a unified approach to statistical language modeling for Chinese.* ACM TALIP, 1(1): 3-33.

Gao, Jianfeng, Mu Li and Changning Huang. 2003. *Improved source-channel models for Chinese word segmentation.* In ACL-2003. Sapporo, Japan.

Gao, Jianfeng, Andi Wu, Mu Li, Chang-Ning Huang, Hongqiao Li, Haowei Qin and Xinsong Xia. 2004. *Adaptive Chinese Word Segmentation.* In proceedings of ACL 2004. Barcelona, Spain.

Ge, X., Pratt, W. and Smyth, P. 1999. *Discovering Chinese Words from Unsegmented Text.* SIGIR-99, pages 271-272.

Hockenmaier, J. and Brew, C. 1998. *Error-driven learning of Chinese word segmentation*. In J. Guo, K. T. Lua, and J. Xu, editors, 12th Pacific Conference on Language and Information, pp. 21̃8229, Singapore. Chinese and Oriental Languages Processing Society.

Katz, S. M. 1987. *Estimation of probabilities from sparse data for the language model component of a speech recognizer.* IEEE Trans. Acoustics, Speech Signal Process. ASSP-35, 3 (March), 400-401.

Palmer, David. 1997. *A Trainable Rule-Based Algorithm for Word Segmentation.* In Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics (ACL '97), Madrid.

Sun, Jian, Jianfeng Gao, Lei Zhang, Ming Zhou, and Changning Huang. 2002. *Chinese named entity identification using class-based language model.* In: COLING 2002. Taipei, Taiwan.

Wu, Andi. 2003. *Chinese Word Segmentation in MSR-NLP.* In Proceedings of the Second SIGHAN Workshop on Chinese Language Processing, Sapporo, Japan.