Automatic Corpus Expansion for Chinese Word Segmentation by Exploiting the Redundancy of Web Information

Xipeng Qiu, ChaoChao Huang and Xuanjing Huang
Shanghai Key Laboratory of Intelligent Information Processing
School of Computer Science, Fudan University, Shanghai, China
xpqiu@fudan.edu.cn, superhuang007@gmail.com, xjhuang@fudan.edu.cn

Abstract

Currently most of state-of-the-art methods for Chinese word segmentation (CWS) are based on supervised learning, which depend on large scale annotated corpus. However, these supervised methods do not work well when we deal with a new different domain without enough annotated corpus. In this paper, we propose a method to automatically expand the training corpus for the out-of-domain texts by exploiting the redundant information on Web. We break up a complex and uncertain segmentation by resorting to Web for an ample supply of relevant easy-to-segment sentences. Then we can pick out some reliable segmented sentences and add them to corpus. With the augmented corpus, we can re-train a better segmenter to resolve the original complex segmentation. The experimental results show that our approach can more effectively and stably improve the performance of CWS. Our method also provides a new viewpoint to enhance the performance of CWS by automatically expanding corpus rather than developing complicated algorithms or features.

1 Introduction

Word segmentation is a fundamental task for Chinese language processing. In recent years, Chinese word segmentation (CWS) has undergone great development. The popular method is to regard word segmentation as a sequence labeling problems (Xue, 2003; Peng et al., 2004). The goal of sequence labeling is to assign labels to all elements in a sequence, which can be handled with supervised learning algorithms, such as Maximum Entropy (ME) (Berger et al., 1996), Conditional Random Fields (CRF)(Lafferty et al., 2001).

After years of intensive researches, Chinese word segmentation achieves a quite high precision. However, the performance of segmentation is not so satisfying for the practical demands to analyze Chinese texts. The key reason is that most of annotated corpora are drawn from news texts. Therefore, the system trained on these corpora cannot work well with the out-of-domain texts.

Since these supervised approaches often has a high requirement on the quality and quantity of annotated corpus, which is always not easy to create. As a result, many methods were proposed to utilize the information of unlabeled data.

There are three kinds of methods for domain adaptation problem in CWS.

The first is to use unsupervised learning algorithm to segment texts, like branching entropy (BE) (Jin and Tanaka-Ishii, 2006), normalized variation of branching entropy (nVBE)(Magistry and Sagot, 2012).

Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, pages 1154–1164, Dublin, Ireland, August 23-29 2014.

The second is to use unsupervised or domain-independent features in supervised learning for Chinese word segmentation, such as punctuation and mutual information(MI), word accessory variance (Feng et al., 2004; Zhao and Kit, 2008; Sun and Xu, 2011)

The third is to use semi-supervised learning (Zhu, 2005) in sequence labeling to address the difference in source and target distributions (Jiao et al., 2006; Altun et al., 2006; Suzuki and Isozaki, 2008).

Although these methods improve the performance of out-of-domain texts, the performance is still worse than that of in-domain texts obviously.

We firstly investigate the reasons of lower performance in new domain for state-of-the-art CWS systems and find that most of error segmentation were caused by out-of-vocabulary (OOV) words, also called new words or unknown words (see details in Section 3). It is difficult to devote efforts to building a corpus for out-of-domain texts, since new words are produced frequently as the development of the society, especially the Internet society. It is also impractical to manually maintain an up-to-date corpus to include all geographical names, person names, organization names, technical terms, etc.

In this paper, we propose a method to automatically expand the training corpus for the out-of-domain texts by exploiting the redundant information on Web. When we meet a complex and potentially difficult-to-segment sentence, we do not expect to solve it with more complicated learning algorithm or elaborate features. We assume that there are some relevant sentences that are relatively easy to process. These simple sentences can help to solve the complex one.

For example, the sentence “欧莱雅美宝莲 (L'Oreal, Maybelline)” is difficult to segment if both “欧莱雅 (L'Oreal)” and “美宝莲 (Maybelline)” are unknown words. However, we can always find some easy-to-segment sentences, such as “我使用美宝莲 (I use Maybelline)”, “欧莱雅的产品 (production of L'Oreal)”, and so on. When we use these simple sentences to re-train the segmenter, we can solve the previous complex sentence.

Our method relies on breaking up the complex problems into relevant smaller, simpler problems that can be solved easily. Fortunately, we can resort to the scale and redundancy of the web for an ample supply of simple sentences that are relatively easy to process.

Our method is very easy to implement upon a trainable base segmenter. Given the out-of-domain texts, we firstly choose some uncertain segmentations and select the candidate expansion seeds. Secondly, we use these seeds to get the relevant texts from Web search engine. Then we segment these texts and add the texts with high confidence to training corpus. Finally, we can get a better segmenter with the new corpus.

The rest of the paper is organized as follows: we review the related works in section 2. In section 3, we analyze the influence factor for CWS. Then we describe our method in section 4. Section 5 introduces the base segmenter. Section 6 gives the experimental results. Finally we conclude our work in section 7.

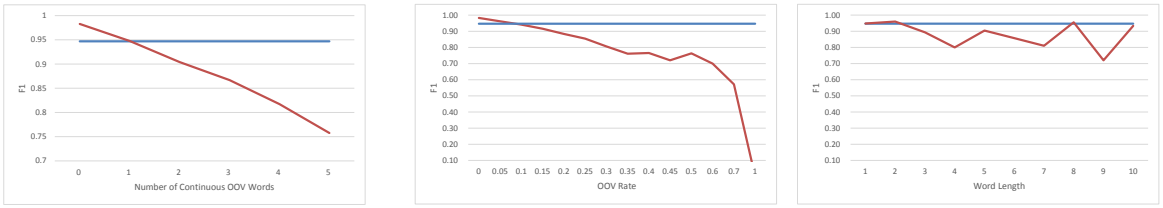
2 Related Works

The idea of exploring information redundancy on Web was introduced in question answering system (Kwok et al., 2001; Clarke et al., 2001; Banko et al., 2002) and the famous information extraction system KNOWITALL(Etzioni et al., 2004). However, this idea is rarely mentioned in Chinese word segmentation.

Nonetheless, there are three kinds of related methods on Chinese word segmentation.

One is active learning. Both (Li et al., 2012) and (Sassano, 2002) try to use active learning method to expand annotated corpus, but they still need to manually label some new raw texts in order to enlarge the training corpus. Different with these methods, our method do not require any manual oracle labeling at all.

Another is self-training, also called bootstrapping or self-teaching (Zhu, 2005). Self-training is a general semi-supervised learning approach. In self-training, a classifier is first trained with the small amount of labeled data. The classifier is then used to classify the unlabeled data.



(a) Number of continuous OOV words (b) OOV rate (c) Word Length

The blue horizontal line is the overall F1 score, and the red line is the F1 scores with different values of the factor.

Figure 1: Analysis of Influence Factors

Typically the most confident unlabeled points, together with their predicted labels, are added to the training set. The classifier is re-trained and the procedure repeated. Note that the classifier uses its own predictions to teach itself. Self-training has been applied to several natural language processing (NLP) tasks, such as word sense disambiguation (Yarowsky, 1995), POS-tagging (Clark et al., 2003; Jiang and Zhai, 2007; Liu and Zhang, 2012), parsing (Steedman et al., 2003; McClosky et al., 2006; Reichart and Rappoport, 2007; Sagae, 2010), information extraction(Etzioni et al., 2004)and so on. It has been proven that self-training can improve system performance on the target domain by simultaneously modeling annotated source-domain data and unannotated target domain data in the training process. However, the data on target domain cannot always help itself (Steedman et al., 2003).

The third is weakly supervised learning. (Li and Sun, 2009; Jiang et al., 2013) utilized the massive manual natural annotations or punctuation information on the Internet to improve the performance of CWS. However, these natural annotations are just partial annotations and their roles depend on the qualities of the selected resource, such as Wikipedia.

In this paper, we wish to propose a method to obtain new fully-annotated data in more aggressive way, which can combine the advantages of the above works.

3 Analysis of Influence Factors for CWS

Before describing our method, we give an analysis of the impact of out-of-vocabulary (OOV) words for segmentation. We first conduct experiments on the Chinese Treebank (CTB6.0) dataset (Xue et al., 2005) (The detailed information of dataset is shown in Section 6).

Table 1 shows the performance of base segmenter. The F1 score of OOV words is significantly lower than that of in-vocabulary (INV) words.

	Precision	Recall	F1
INV	95.86	96.58	96.21
OOV	74.12	66.77	70.25
Total	94.64	94.73	94.69

Table 1: Performances of INV and OOV words

We also investigate the impacts of three different factors: number of continuous OOV words, OOV rate and word length. Figure 1 shows the F1 scores with the changes of the different factors. We find that OOV words significantly improve the difficulty of segmentation, while the word length does not always harm the accuracy.

These findings also indicate that we can improve the performance of CWS if we have a dictionary or annotated corpus including these OOV words. With the redundancy of the Web information, it is not difficult to automatically obtain the expected dictionary or corpus.

4 Our Method

In this section, we describe our method to automatically expand the training corpus.

4.1 Framework of Automatic Corpus Expansion

Our framework of automatic corpus expansion is similar to standard process self-training or active learning for domain adaptation. Given a trainable base segmenter, the texts in out-of-domain, we firstly choose some uncertain segmentations and select the candidate expansion seeds. Secondly, we use these seeds to get the relevant texts from Web search engine. Then we segment these texts and add the texts with high confidence to training corpus. Finally, we can get a better segmenter with the new corpus.

Algorithm 1 illustrates the framework of automatic corpus expansion.

Algorithm 1 Framework of Automatic Corpus Expansion

Input:
Annotated Corpus C_A
Unannotated Corpus in Target domain C_T
Uncertainty Threshold T_u
Seed Extraction Threshold T_{se}
Acceptation Threshold T_a
Maximum Iteration Number: M

Output: Expanded Annotated Corpus C_A

1: **for** $i = 1$ to M **do**
2: Train a basic segmenter using current C_A with base learner
3: Use the basic segmenter to do segmentation for each sentence in C_T and calculate its confidence.
4: Choose out the sentences collection C_{TS}, in which the segmentation confidence of each sentence is less than T_u.
5: Extract the expansion seeds collection C_{seeds} from C_{TS} and use search engine to acquire relevant raw texts C_{RRT}.
6: Segment and calculate the confidence for each sentence in C_{RRT}.
7: Pick the reliable segmentations C_{new} with confidence more than T_a from C_{RRT}.
8: Add C_{new} into C_A.
9: **end for**
10: **return** C_A;

4.2 Uncertainty Sampling

The first key step in our method is to find the uncertain segmentations. There are many proposed uncertainty measures in the literature of active learning (Settles, 2010), such as entropy and query-by-committee (QBC) algorithm.

In our works, we investigate four following uncertainty measures for each sentence x. We use $S_1(x), S_2(x), \cdots, S_N(x)$ to represent the top N scores given by the segmenter.

Normalized Score U_{NS}

The first measures is normalized score by the length of x, the normalized score U_{NS} is calculated by

$$U_{NS} = \frac{S_1(x)}{L} \tag{1}$$

where L is the length of x.

Standard Deviation U_{SD}

The standard deviation is calculate with the top N scores.

$$U_{SD} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(S_i(x) - \mu)^2} \tag{2}$$

where $\mu = \frac{1}{N}\sum_{i=1}^{N} S_i(x)$ is the average or expected value of $S_i(x)$.

Entropy $U_{Entropy}$

Entropy is a measure of unpredictability or information content. Since we use character-based method for word segmentation, each character is labeled as one of {**B**, **M**, **E**, **S**} to indicate the segmentation. {**B**, **M**, **E**} represent *Begin*, *Middle*, *End* of a multi-character segmentation respectively, and **S** represents a *Single* character segmentation.

Given the top N labeled results for a sentence, each labeled sequence consists of the labels {**B**, **M**, **E**, **S**}. We define $l \in \{B, M, E, S\}$ to represent the label variable, and $\mathbf{count}_j(l)$ to be the number of occurrences of l on position j among the top N results. Thus, we can calculate the entropy for the labeling uncertainty of each character.

The entropy $H_j(l)$ for the character on position j is calculated by

$$H_j(l) = -\sum_{l} \frac{\mathbf{count}_j(l)}{N} \log \frac{\mathbf{count}_j(l)}{N}, \tag{3}$$

where $\sum_l \mathbf{count}_j(l) = N$.

The entropy of sentence $U_{Entropy}$ is the sum of the entropies of all the characters in the sentence.

$$U_{Entropy} = \sum_{j=1}^{L} H_j(l). \tag{4}$$

Margin U_{Margin}

Margin is the deviation of top 2 scores, which is often used in machine learning algorithms, such as support vector machine (Cristianini and Shawe-Taylor, 2000) and passive-aggressive algorithm (Crammer et al., 2006).

$$U_{Margin} = S_1(x) - S_2(x) \tag{5}$$

Among the above four measures, the larger the entropy is, the more uncertain the result is. For the rest three factors, the less the score is, the more uncertain the result is.

We test these four uncertainty measures on the development set in order to choose the best one as our confidence measure.

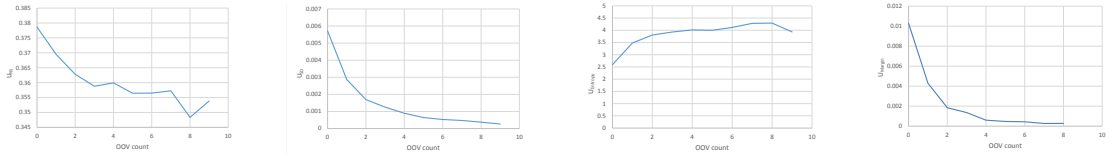
In figure 2, we illustrate the relationship between each uncertainty measure and the OOV count. We assume that the more OOV words are, the more uncertainty is. Meanwhile, a steep learning curve imply a good ability to distinguish whether the result is uncertain.

Obviously, the entropy is not helpful according to our assumption. The normalized score is okay but not good, and both the standard deviation and margin seem to be useful because they can give a better threshold to distinguish uncertain segmentation. Finally, we choose margin as our uncertainty measure.

4.3 Expansion Seeds Extraction

For the uncertain segmentation, not every word is unreliable. We just pick the suspicious fragments. Therefore, we need to extract some seed phrases to get the relevant texts. It is notable that these seed phrases do not need to be words. They can be the combinations of several words or only parts of words.

Take the following sentence for example.



(a) NScore (b) STDEV (c) Entropy (d) Margin

Figure 2: Different Uncertainty Measures

欧莱雅美宝莲兰蔻是很好的品牌
(L'Oreal, Maybelline, Lancome are good brands)

The first fragment "欧莱雅美宝莲兰蔻" is difficult to segment if these words does not appear in training corpus. Conversely, the second fragment is easy to segment since the containing words are very common.

We use base segmenter to get the top five results as follows:

	欧	莱	雅	美	宝	莲	兰	蔻	是	很	好	的	品	牌
1	B	M	M	M	E	B	M	E	S	B	E	S	B	E
2	B	M	M	E	B	E	B	E	S	B	E	S	B	E
3	B	M	E	B	E	B	M	E	S	B	E	S	B	E
4	S	B	M	M	E	B	E	S	S	B	E	S	B	E
5	S	B	M	M	M	E	B	E	S	B	E	S	B	E

(Li et al., 2012) proposed a good way to select the candidate words for active learning with diversity measurement to avoid duplicate annotation. However, their method is not suitable for our work. The reason is that they regarded CWS as a binary classification problem, while our base segmenter uses 1st-order sequence labeling.

In our work, we choose the expansion seeds by calculating the entropy of each character. If the entropy of the character is larger than threshold T_{se}, we say that this character may be in an uncertain context. Thus, we extract the consecutive uncertain characters and their contexts as the expansion seeds.

For the above example, we select the "欧莱雅美宝莲兰蔻 (L'Oreal, Maybelline, Lancome)" and its context "是 (is)" as a seed "欧莱雅美宝莲兰蔻是".

4.4 Collect relevance texts by using Web Search Engines

After obtaining the expansion seeds, we collect the relevant texts on multiple search engines including Google, Baidu and Bing.

For the seed "欧莱雅美宝莲兰蔻是", we can get the following relevance sentence, which is easy to segment.

欧莱雅拥有兰蔻、欧莱雅、美宝莲、薇姿等 500 多个品牌
(L'Oreal owns more than 500 brands, including Lancome, L'Oreal, Maybelline, Vichy, etc.)

In our work, we just get the top 100 relevant texts returned by each search engine without manual intervention. We do not use any search API and directly use the returned webpages by search engine, then extract the snippets and titles. Therefore, we just write a simple program to collect the webpages and clean them.

4.5 Expand Training Corpus

Since the qualities of these relevant texts are spotty, we just pick the reliable texts with high confidence scores. In contrast to uncertainty sampling, we find the certain segmentations from the collecting raw texts and add them to training corpus. Here, we also use a margin to find the reliable ones as new training data.

In our experiments, the number of selected sentence is $1 \sim 5$ for each seed.

Thus, we can re-train a new segmenter on the expanded corpus. After several iteration, we will get a segmenter with the best performance.

5 Base Segmenter

We use discriminative character-based sequence labeling for base word segmentation. Each character is labeled as one of {B, M, E, S} to indicate the segmentation.

We use online Passive-Aggressive (PA) algorithm (Crammer and Singer, 2003; Crammer et al., 2006) to train the model parameters. Following (Collins, 2002), the average strategy is used to avoid the overfitting problem.

6 Experiment

To evaluate our algorithm, we use both CTB6.0 and CTB7.0 datasets in our experiments. CTB is a segmented, part-of-speech tagged, and fully bracketed corpus in the constituency formalism. It is also a popular data set to evaluate word segmentation methods, such as (Sun and Xu, 2011). Since CTB dataset is collected from different sources, such as newswire, magazine, broadcast news and web blogs, it is suitable to evaluate the performance of CWS systems on different domains.

We conduct two experiments on different divisions of datasets.

1. The first experiment is performed on CTB6.0 for comparison with state-of-the-art systems which also utilize the unlabeled data for word segmentation.

2. The second experiment is performed on CTB7.0 for better evaluation on out-of-domain texts. CTB7.0 contains some newer news texts and web blogs texts, which is more suitable to evaluate our method for out-of-domain data.

In our experiments, we set $\mathcal{C} = 0.01$ for PA algorithm. We also try to use the different values of $\mathcal{C}$, and found that larger values of $\mathcal{C}$ imply a more aggressive update step and result to fast convergence, but it has little influence on the final accuracy. The maximum iteration number M' of PA algorithm is set to 50.

The feature templates are $C_iT_0, (i = -1, 0, 1)$,$C_{-1,0}T_0$, $C_{0,1}T_0, C_{-1,1}T_0$, $T_{-1,0}$. C represents a Chinese character, and the subscript of C indicates its position relative to the current character, whose subscript is 0. T represents the character-based tag.

The evaluation measure are reported are precision, recall, and an evenly-weighted F_1.

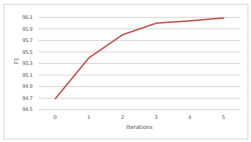
6.1 Experiments on CTB6.0

Train	Dev	Test
81-325, 400-454, 500-554, 590-596, 600-885, 900, 1001-1017, 1019, 1021-1035, 1037-1043, 1045-1059, 1062-1071, 1073-1078, 1100-1117, 1130-1131 1133-1140, 1143-1147, 1149-1151,2000-2139, 2160-2164, 2181-2279,2311-2549, 2603-2774, 2820-3079	41-80, 1120-1129, 2140-2159, 2280-2294, 2550-2569, 2775-2799, 3080-3109	(1-40,901-931 newswire) (1018, 1020, 1036, 1044,1060-1061, 1072, 1118-1119, 1132,1141-1142, 1148 magazine) (2165-2180, 2295-2310, 2570-2602, 2800-2819, 3110-3145 broadcast news)

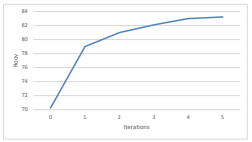
Table 2: CTB6.0 Dataset Division

On CTB 6.0, we divide the training, development and test sets according to (Yang and Xue, 2012). , which are shown in Table 2 The detailed statistical information is shown in Table 3.

Firstly, We use the development set to determine the parameters in Algorithm 1. For T_u, T_{se} and T_a, we have three rounds to determine the parameters. In first round, we find the best value $t1$ in the range to $0 \sim 1$ with the interval of 0.1. In second round, we find the best value $t2$ in range $t1 - 0.1 \sim t1 + 0.1$ with the interval of 0.01. In third round, we find the final best value $t3$

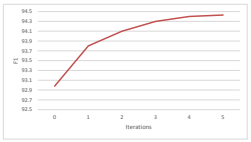


(a) F1 score

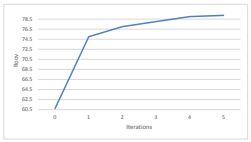


(b) OOV Recall

Figure 3: Iterative Learning Curve on CTB6.0



(a) F1 score



(b) OOV Recall

Figure 4: Iterative Learning Curve on CTB7.0

in the range to $t2 - 0.01 \sim t1 + 0.01$ with the interval of 0.001. The maximum iteration number M is just determined based on convergence with the range $1 \sim 10$.

Finally, we set these parameters as following: uncertainty threshold $T_u = 0.003$, seed extraction threshold $T_{se} = 0.65$, acceptation threshold $T_a = 0.004$ and maximum iteration number $M = 5$.

Figure 3 shows the changing curve of F1 and OOV recall in the process of corpus expansion. The performance of the baseline segmenter is shown at iteration 0. The curve shows that the F1 score and OOV recall have continuous improvement with the increasing of train corpus. The maximum performance is achieved at the $5th$ iteration. The detailed results are shown in Table 4. Compared with the baseline, the expanded corpus leads to a segmenter with significantly higher accuracy. The relative error reductions are 26.37% and 43.63% in terms of the balanced F-score and the recall of OOV words respectively.

Dataset	Sents	Words	Chars	OOV Rate
Train.	22757	639506	1053426	-
Dev.	2003	59764	100038	5.45%
Test	2694	81304	133798	5.58%

Table 3: Corpus Information of CTB 6.0

Test	P	R	F1	R_{oov}
Baseline	94.64	94.73	94.69	70.25
Final	95.66	96.51	96.09	83.23
(Sun and Xu, 2011)	95.86	95.62	95.74	79.28

Table 4: Performance on CTB6.0

6.2 Experiments on CTB7.0

CTB7.0 includes documents from newswire, magazine articles, broadcast news, broadcast conversations, newsgroups and weblogs. The newly added documents contains texts from web blogs, which is very different with news texts. Therefore, we use the documents (No. 4198 4411, weblogs) as test dataset, and the rest as training dataset. The detailed statistical information is shown in Table 5. We can see that the OOV rate is higher than the dataset in the first experiment.

Dataset	Sents	Words	Chars	OOV Rate
Train.	40425	987307	1601142	-
Test	10177	209827	342061	7.09%

Table 5: Corpus Information of CTB 7.0

Test	P	R	F1	R_{oov}
Baseline	93.58	92.40	92.98	60.72
Final	94.47	94.40	94.43	79.24

Table 6: Performance on CTB7.0

Figure 4 shows the changing curve of F1 and OOV recall in the process of corpus expansion. The performance of the baseline segmenter is shown at iteration 0. The curve shows that the F1 score and OOV recall have continuous improvement with the increasing of train corpus. The maximum performance is achieved at iteration 5. The detailed results are shown in Table 6. Compared with the baseline, the expanded corpus leads to a segmenter with significantly higher accuracy. The relative error reductions are 20.66% and 47.15% in terms of the balanced F-score and the recall of OOV words respectively.

6.3 Analysis

The experimental results show that our method is very effective to improve the performance of Chinese word segmentation. Especially, our method gives a significant boost on OOV words.

For the words such as "门兴格拉德巴赫 (Borussia Moenchengladbach)", "过氧化氢酶 (catalase)", "易中天 (Yi ZhongTian, a Chinese person name)" and "黄金档 (prime time)", it is still difficult to segment them correctly even if we can obtain useful features from unlabeled data. When we take advantage of the redundant information from Web, we can easily collect the relevant easy-to-segment sentences to expand the training corpus.

Our method can result to a segmenter significantly better than the systems which finds the informative features derived from unlabeled data, such as (Sun and Xu, 2011). This also suggests that expanding corpus is more effective than developing complicated algorithm or well-design features. Of course, our method is compatible with these technologies, which can further improve the performance of CWS by combining the Web redundancy.

7 Conclusion

In this paper, we propose a method to automatically expand the training corpus for the out-of-domain texts. Given the out-of-domain texts, we first choose some uncertain segmentations as candidate expansion seeds, and use these seeds to get the relevant texts from search engine. Then we segment the texts and add the texts with high confidence to training corpus. We can always obtain some easily-segmented texts due to the large amount of redundancy texts on Web, especially for new words. Our experimental results show that our proposed method can more effectively and stably utilize the unlabeled examples to improve the performance. Our method also provides a new viewpoint to enhance the performance of CWS by expanding corpus rather than developing complicated algorithms or features.

The long term goal of our method is to build an online and constant learning system, which can identify the difficult tasks and seek help from crowdsourcing. Search engines are special cases of crowdsourcing. In the future, we wish to investigate our method for other NLP tasks, such as POS tagging, Named Entity Recognition, and so on.

Acknowledgments

We would like to thank the anonymous reviewers for their valuable comments. This work was funded by NSFC (No.61003091), Science and Technology Commission of Shanghai Municipality (14ZR1403200) and Shanghai Leading Academic Discipline Project (B114).

References

Y. Altun, D. McAllester, and M. Belkin. 2006. Maximum margin semi-supervised learning for structured variables. *Advances in neural information processing systems*, 18:33.

Michele Banko, Eric Brill, Susan Dumais, and Jimmy Lin. 2002. AskMSR: Question answering using the worldwide web. In *Proceedings of 2002 AAAI Spring Symposium on Mining Answers from Texts and Knowledge Bases*, pages 7–9.

A.L. Berger, V.J. Della Pietra, and S.A. Della Pietra. 1996. A maximum entropy approach to natural language processing. *Computational Linguistics*, 22(1):39–71.

Stephen Clark, James R Curran, and Miles Osborne. 2003. Bootstrapping POS taggers using unlabelled data. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*, pages 49–55. Association for Computational Linguistics.

C.L.A. Clarke, G.V. Cormack, and T.R. Lynam. 2001. Exploiting redundancy in question answering. *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 358–365.

Michael Collins. 2002. Discriminative training methods for hidden markov models: Theory and experiments with perceptron algorithms. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing*.

K. Crammer and Y. Singer. 2003. Ultraconservative online algorithms for multiclass problems. *Journal of Machine Learning Research*, 3:951–991.

Koby Crammer, Ofer Dekel, Joseph Keshet, Shai Shalev-Shwartz, and Yoram Singer. 2006. Online passive-aggressive algorithms. *Journal of Machine Learning Research*, 7:551–585.

N. Cristianini and J. Shawe-Taylor. 2000. *An introduction to support Vector Machines: and other kernel-based learning methods*. Cambridge Univ Pr.

Oren Etzioni, Michael Cafarella, Doug Downey, Stanley Kok, Ana-Maria Popescu, Tal Shaked, Stephen Soderland, Daniel S. Weld, and Alexander Yates. 2004. Web-scale information extraction in knowitall: (preliminary results). In *Proceedings of the 13th international conference on World Wide Web*, WWW '04, pages 100–110, New York, NY, USA. ACM.

H. Feng, K. Chen, X. Deng, and W. Zheng. 2004. Accessor variety criteria for chinese word extraction. *Computational Linguistics*, 30(1):75–93.

Jing Jiang and ChengXiang Zhai. 2007. Instance weighting for domain adaptation in NLP. In *ACL*, volume 2007, page 22.

Wenbin Jiang, Meng Sun, Yajuan Lü, Yating Yang, and Qun Liu. 2013. Discriminative learning with natural annotations: Word segmentation as a case study. In *ACL*, pages 761–769.

Feng Jiao, Shaojun Wang, Chi-Hoon Lee, Russell Greiner, and Dale Schuurmans. 2006. Semi-supervised conditional random fields for improved sequence segmentation and labeling. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 209–216. Association for Computational Linguistics.

Zhihui Jin and Kumiko Tanaka-Ishii. 2006. Unsupervised segmentation of Chinese text by use of branching entropy. In *Proceedings of the COLING/ACL on Main conference poster sessions*, pages 428–435. Association for Computational Linguistics.

C.C.T. Kwok, O. Etzioni, and D.S. Weld. 2001. Scaling question answering to the web. *Proceedings of the 10th international conference on World Wide Web*, pages 150–161.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*.

Zhongguo Li and Maosong Sun. 2009. Punctuation as implicit annotations for chinese word segmentation. *Computational Linguistics*, 35(4):505–512.

Shoushan Li, Guodong Zhou, and Chu-Ren Huang. 2012. Active learning for Chinese word segmentation. In *COLING (Posters)*, pages 683–692.

Yang Liu and Yue Zhang. 2012. Unsupervised domain adaptation for joint segmentation and pos-tagging. In *COLING (Posters)*, pages 745–754.

Pierre Magistry and Benoît Sagot. 2012. Unsupervized word segmentation: the case for mandarin chinese. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2*, pages 383–387. Association for Computational Linguistics.

David McClosky, Eugene Charniak, and Mark Johnson. 2006. Effective self-training for parsing. In *Proceedings of the main conference on human language technology conference of the North American Chapter of the Association of Computational Linguistics*, pages 152–159. Association for Computational Linguistics.

F. Peng, F. Feng, and A. McCallum. 2004. Chinese segmentation and new word detection using conditional random fields. *Proceedings of the 20th international conference on Computational Linguistics*.

Roi Reichart and Ari Rappoport. 2007. Self-training for enhancement and domain adaptation of statistical parsers trained on small datasets. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 616–623, Prague, Czech Republic, June. Association for Computational Linguistics.

Kenji Sagae. 2010. Self-training without reranking for parser domain adaptation and its impact on semantic role labeling. In *Proceedings of the 2010 Workshop on Domain Adaptation for Natural Language Processing*, pages 37–44. Association for Computational Linguistics.

Manabu Sassano. 2002. An empirical study of active learning with support vector machines for japanese word segmentation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 505–512. Association for Computational Linguistics.

Burr Settles. 2010. Active learning literature survey. *University of Wisconsin, Madison.*

Mark Steedman, Miles Osborne, Anoop Sarkar, Stephen Clark, Rebecca Hwa, Julia Hockenmaier, Paul Ruhlen, Steven Baker, and Jeremiah Crim. 2003. Bootstrapping statistical parsers from small datasets. In *Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics-Volume 1*, pages 331–338. Association for Computational Linguistics.

Weiwei Sun and Jia Xu. 2011. Enhancing Chinese word segmentation using unlabeled data. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 970–979. Association for Computational Linguistics.

Jun Suzuki and Hideki Isozaki. 2008. Semi-supervised sequential labeling and segmentation using giga-word scale unlabeled data. In *ACL*, pages 665–673. Citeseer.

Naiwen Xue, Fei Xia, Fu-Dong Chiou, and Martha Palmer. 2005. The Penn Chinese TreeBank: Phrase structure annotation of a large corpus. *Natural language engineering*, 11(2):207–238.

Nianwen Xue. 2003. Chinese word segmentation as character tagging. *Computational Linguistics and Chinese Language Processing*, 8(1):29–48.

Yaqin Yang and Nianwen Xue. 2012. Chinese comma disambiguation for discourse analysis. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 786–794. Association for Computational Linguistics.

D. Yarowsky. 1995. Unsupervised word sense disambiguation rivaling supervised methods. In *Proceedings of the 33rd annual meeting on Association for Computational Linguistics*, pages 189–196. Association for Computational Linguistics.

H. Zhao and C. Kit. 2008. Unsupervised segmentation helps supervised learning of character tagging for word segmentation and named entity recognition. In *The Sixth SIGHAN Workshop on Chinese Language Processing*, pages 106–111. Citeseer.

Xiaojin Zhu. 2005. Semi-supervised learning literature survey. Technical Report 1530, Computer Sciences, University of Wisconsin-Madison.