Active Learning for Chinese Word Segmentation

Shoushan Li[1] *Guodong Zhou*[1] *Chu-Ren Huang*[2]
(1) Natural Language Processing Lab, Soochow University, China
(2) Department of Chinese and Bilingual Studies, The Hong Kong Polytechnic University, Hong Kong
{lishoushan, gdzhou}@suda.edu.cn, churenhuang@gmail.com

ABSTRACT

Currently, the best performing models for Chinese word segmentation (CWS) are extremely resource intensive in terms of annotation data quantity. One promising solution to minimize the cost of data acquisition is active learning, which aims to actively select the most useful instances to annotate for learning. Active learning on CWS, however, remains challenging due to its inherent nature. In this paper, we propose a Word Boundary Annotation (WBA) model to make effective active learning on CWS possible. This is achieved by annotating only those uncertain boundaries. In this way, the manual annotation cost is largely reduced, compared to annotating the whole character sequence. To further minimize the annotation effort, a diversity measurement among the instances is considered to avoid duplicate annotation. Experimental results show that employing the WBA model and the diversity measurement into active learning on CWS can save much annotation cost with little loss in the performance.

KEYWORDS: Chinese Word Segmentation; Active Learning; Word Boundary Annotation

Proceedings of COLING 2012: Posters, pages 683–692,
COLING 2012, Mumbai, December 2012.

1 Introduction

Chinese word segmentation (CWS) is an indispensable pre-processing requirement for many Chinese language processing tasks, such as named entity recognition, syntactic parsing, semantic parsing, information extraction, and machine translation. Although state-of-the-art CWS systems report a high performance at the level of 95-97%, these systems typically require a large scale of pre-segmented corpus of tens (if not hundreds) of millions of words for training. However, the collection of the data on such a scale is very time-consuming and resource-intensive.

One possible solution to handle this dilemma is to deploy active learning, where only a small scale of instances are actively selected to serve as training data so that the annotation effort can be highly reduced (Settles and Craven, 2008). Although active learning has been widely employed to many NLP tasks, such as word sense disambiguation (Chan and Ng, 2007; Chen et al., 2006; Fujii et al., 1998), text categorization (Lewis and Gale, 1994; Liere and Tadepalli, 1997; McCallum and Nigam, 1998; Li et al., 2012), and named entity recognition (Shen et al., 2004), there are few studies of active learning on CWS, probably due to the strong challenges inherent in performing active learning on CWS.

First, the state-of-the-art methods treat CWS as a sequence labelling task (Jiang et al., 2008; Ng and Low, 2004; Tseng et al., 2005; Zhang et al., 2006), i.e. labelling characters with tags from a pre-defined tag set, representing the position of a character in a word. Different from traditional classification tasks, each character is tagged sequentially according to its corresponding context. Under this circumstance, a character cannot be determined as a single unit to query in active learning. One possible solution is to select one sentence as a unit for annotation, as Sassano (2002) does for Japanese word segmentation. However, such solution is expensive for annotation and since one sentence might contain some words which can be easily segmented correctly by existing models with high confidence, annotating them becomes a waste of time and manual effort.

Second, the number of the characters in a CWS corpus is normally extremely huge. For example, among the four corpora in SIGHAN Bakeoff 2 (Emerson, 2005), even the smallest corpus contains more than 1,800,000 characters while others are much larger in the order of tens of millions of characters. Compared to other tasks like text classification, normally with less than 20,000 instances (McCallum and Nigam, 1998), or named entity recognition, normally with less than 80,000 instances (Shen et al., 2004), CWS with such tremendous amount of instances makes it impossible to iteratively select one most informative instance for manual annotation in the active learning process. Instead, in each iteration, many informative instances are selected at the same time in practice. Under this circumstance, the selected informative instances are very likely overlapping when a standard uncertainty query strategy is used. For example, one unknown word may appear many times and a few sentences containing the unknown word may be selected for manual annotation at the same time according to the uncertainty strategy.

In this paper, we address the above challenges in active learning for CWS. In particular, for the first challenge, we propose a word boundary annotation (WBA) model, where the boundary between a character pair is considered the annotation unit. Specifically, we actively select the most informative boundaries to label manually and leave their easy and non-informative surrounding boundaries automatically labelled. Compared to using the sentence as the annotation unit, using the boundary is capable of reducing much annotation cost. For the second challenge, we propose

a diversity measurement among the instances to avoid duplicate annotation, so as to further reduce the annotation efforts.

2 Related Work

Research on CWS has a long history and various methods have been proposed in the literature. Basically, these methods are mainly focus on two categories: unsupervised and supervised.

Unsupervised methods aim to build a segmentation system without any lexicon or labelled data. They often start from an empirical definition of a word and then use some statistical measures, e.g. mutual information (Sproat and Shih, 1990; Sun et al., 1998), to learn words from a large unlabelled data resource. Although these unsupervised methods can capture many strong words, their performance is often not high enough for the practical use.

Supervised methods, such as HMM tagging (Xue, 2003), character-based classification (Wang et al., 2008) and morpheme-based lexical chunking (Fu et al., 2008), attempt to acquire a model based on a dictionary or a labelled data set. Among them, character-based classification has drawn most attention recently and been further implemented with sequence labelling algorithms (Tseng et al., 2005), e.g., conditional random fields (CRF), which perform well in both in-vocabulary (IV) recall and out-of-vocabulary (OOV) recall. Based on the character labelling approach, many related studies make efforts to improve the performance by various means, such as using more tags and features (Tang et al., 2009; Zhao et al., 2006), employing word-based tagging without tagging (Zhang and Clark, 2007), employing some joint models that combines a generative model and a discriminative model (Wang et al., 2010; Wang et al., 2011) or Markov and semi-Markov CRF (Andrew, 2006), and integrating unsupervised segmentation features (Zhao and Kit, 2011).

Although there are various studies CWS individually, there are few studies of active learning on CWS. One related work is about active learning on Japanese word segmentation via Support Vector Machines (SVM) (Sassano, 2002). However, both the two challenging problems mentioned above are unsolved. Specifically, that study annotates the whole sentence as a basic unit, which means much more annotation effort than our model. Furthermore, our corpus scale is much larger than the one in Sassano (2002). This makes SVM impractical in terms of the training time for active learning on CWS. Meanwhile, they do not give an explicit diversity measurement, although their two-pool strategy implicitly considers the diversity.

3 Our Approach

3.1 Framework of Active Learning for CWS

Figure 1 illustrates the framework of our active learning approach for CWS. In the following subsections, we address the two remaining key issues.

1) The Word Boundary Annotation (WBA) model, which cares boundary annotation instead of the whole sentence.

2) The sample selection strategy $\phi(x)$, which evaluates the informativeness of one instance x. An efficient selection strategy is essential for active learning on CWS, where a huge number of unlabeled instances are involved.

Input:
Labeled set *L*, unlabeled pool *U*, selection strategy $\phi(x)$

Procedure:
Repeat until the predefined stopping criterion is met
(1). Learn a segmenter using current *L* with WBA
(2). Use current segmenter to label all the unlabeled boundaries
(3). Use the selection strategy $\phi(x)$ to select a batch of most informative boundaries for oracle labelling
(4). Put the new labeled boundaries together with their context (automatically labeled) into *L*

Figure 1: WBA-based active learning for CWS

3.2 Word Boundary Annotation (WBA) Model

3.2.1 Boundary Labelling

Formally, a Chinese text can be formalized as a sequence of characters and intervals

$$c_1 I_1 c_2 I_2, ..., c_{n-1} I_{n-1} c_n$$

where c_i means a character and I_i means an interval between two characters. Since there is no indication of word boundaries in a Chinese text, each interval might be a word boundary ($I_i = 1$) or not ($I_i = 0$). Accordingly, the objective of manual annotation is to label the word boundaries given the sequence of characters.

Take following sentence **E-A** as an example, where '/' in the output indicates a word boundary. The annotation process is to indicate that the intervals of I_{A3}, I_{A5}, I_{A7}, I_{A8}, I_{A10}, I_{A12}, I_{A13}, I_{A16}, I_{A18}, and I_{A19}.

E-A. Input: 索 I_{A1} 拉 I_{A2} 纳 I_{A3} 今 I_{A4} 天 I_{A5} 下 I_{A6} 午 I_{A7} 在 I_{A8} 波 I_{A9} 兰 I_{A10} 议 I_{A11} 会 I_{A12} 上 I_{A13} 发 I_{A14} 表 I_{A15} 了 I_{A16} 演 I_{A17} 说 I_{A18} 。 I_{A19}
Output: 索拉纳/ 今天/ 下午/ 在/ 波兰/ 议会/ 上/ 发表了/ 演说/ 。 /
(*Solana gave a speech in the Polish parliament this afternoon .*)

From the above example, we can see that the annotation cost of CWS is very high because too many of boundaries (samples) need to be manually labeled. To overcome this problem, our active learning strategy labels those informative boundaries only.

3.2.2 Context Collection

In the training phase, the context of a selected boundary is essential for learning in that the nearby boundary categories are required to obtain the transition features. Consequently, not only the most informative boundaries but also their surrounding characters and boundaries are required to be collected for generating the new training data. In this paper, the nearby boundaries are automatically determined via the basic segmenter and don't need manual annotation.

In our approach, the context of a manually labelled boundary is defined as the character sequence between the first previous word boundary and the first following word boundary. In particular, if

the selected boundary is manually labelled as a word boundary, i.e. $y_k = 1$, the two words around it are considered as its context. For examples, in the example sentence **E-A**, I_{A1}, I_{A2}, I_{A3} and I_{A9} are among the most informative boundaries. Since I_{A3} is manually labelled as a word boundary, '索拉纳/今天/' are considered as its context with I_{A4} and I_{A5} automatically labelled. In contrast, if the selected boundary is not manually labelled as a word boundary, i.e. $y_k = 0$, only the word containing the selected boundary is considered as its context. For example, I_{A9} is not manually labelled as word boundary and thus only '波兰/' is considered as its context with I_{A8} and I_{A10} automatically labelled.

3.3 Sample Selection Strategy with Diversity Measurement

In the literature, uncertainty sampling (Lewis and Gale, 1994) and Query-By-Committee (QBC) (Seung et al., 1992) are two popular selection schemes in active learning. This paper focuses on uncertainty sampling.

In uncertainty sampling, a learner queries the instance which is most uncertain to label. As WBA is a binary classification problem, uncertainty can simply be measured by querying the boundary whose posterior probability is nearest to 0.5. Therefore, we can define the uncertainty confidence value as follows:

$$\phi^{Un}(b_k) = \max_{y \in \{0,1\}} P(y \mid I_k) - 0.5$$

where $P(y \mid I_k)$ denotes the posterior probability that boundary I_k is labelled as y. The lower the confidence value is, the more informative the boundary is thought to be. After computing the confidences, all the boundaries in the unlabeled pool U are ranked according to their uncertainty values. In this way, a batch of top uncertain boundaries can be picked as the most informative ones for oracle labelling.

A major problem with uncertainty sampling is that it may cause duplicate annotation. That is to say, some instances in the "N-best" queries may be similar. To minimize the manual annotation effort, some diversity measurement among the instances should be taken into account to avoid duplicate annotation. For example, in the example **E-A** above, both the words '索拉纳' and '波兰' are unknown words for the initial segmenter learned by the initial labelled set L with the boundaries of I_{A1}, I_{A2}, I_{A9}, I_{B1}, I_{B2}, and I_{B9}, among the top uncertain instances. Obviously, some boundaries share the same segmentation information, e.g., I_{A1} and I_{B1}. Therefore, labelling both of them is a waste.

One straightforward way to handle such duplicate annotation is to compute the similarity between every two instances and then pick those with the highest diversities (Settles and Craven, 2008). This method, however, requires $O(N^2)$ in computational complexity where N is the number of all boundaries. When N is huge (e.g. N>1,800,000 in our experiments), the high computational burden is simply unacceptable. Fortunately, we find that the similarity between two boundaries is highly related to their surrounding character N-grams (in particular bigrams) and we can better evaluate the diversity with the help of the surrounding character bigrams.

This is done in this paper by recording the frequencies of all surrounding bigrams in a set S_{cc}, where $f_{c_i c_{i+1}} \in S_{cc}$ indicates the frequency of the character bigram $c_i c_{i+1}$ and is initialized to 0.

During training, we go through all the boundaries in the unlabeled data only once and the frequency of the surrounding bigram is updated serially as:

$$f_{c_k c_{k+1}} += 1$$

Where $c_k c_{k+1}$ is the surrounding character bigram of current boundary I_k. Meanwhile, the diversity of boundary I_k can be measured exactly by the frequency of its surrounding bigram:

$$\phi^{Div}(I_k) = f_{c_k c_{k+1}}$$

It is worth mentioning that above diversity measure is a dynamic one. It is possible that two boundaries with the same character bigram context, e.g., I_{A1} and I_{B1} in the above examples, are assigned with different diversity values during training. Specifically, the boundary with a first appearing bigram has the lowest diversity value while the boundaries appearing afterwards have higher values and thus are not likely to be picked as the top informative ones. In this way, the duplicate-annotated words can be avoided to some extent.

In summary, uncertainty sampling with diversity (in short, uncertainty-diversity sampling) ranks the boundaries according to the following formula:

$$\phi^{Un\_Div}(I_k) = \phi^{Un}(I_k) \cdot \phi^{Div}(I_k)$$

The lower the value is, the more informative the boundary is thought to be. Obviously, uncertainty-diversity sampling requires only $O(N)$ in computational complexity.

Therefore, active learning on CWS can be implemented in the following two ways: **Uncertainty sampling**: In each iteration, all the instances in the unlabeled data *U* are ranked according to their uncertainty values and top instances are selected for oracle labelling; **Uncertainty-Diversity sampling**: In each iteration, all the instances in the unlabeled data *U* are ranked according to their uncertainty-diversity values and top instances are selected for oracle labeling.

4 Experimentation

4.1 Experimental Setting

The SIGHAN Bakeoff 2 dataset consists of four different corpora: PKU, MSR, CityU, and AS. But we only report the performance on three of the corpora except AS due to its significant large scale in causing the out-of-memory error. The basic segmenter in the active learning process is trained with a 2-tag labelling model (Huang et al., 2007; Huang and Xue, 2012) and implemented with a public tool for CRF implementation, i.e. CRF++ (Kudo, 2005). For the feature template, we adopt the one by Li and Huang (2009). In all experiments, we use the standard F1 score as our main performance measurement. Besides, the out-of-vocabulary (OOV) recall is used to evaluate the OOV issue.

4.2 Experimental Results

In this experiment, we compare the random selection strategy and the two sampling strategies as illustrated in Section 3.3: uncertainty sampling and uncertainty-diversity sampling. To fairly compare the performances of different sampling strategies, we make sure that the number of annotated boundaries in either uncertainty sampling or uncertainty-diversity sampling is the same as random selection. Figure 2 indicates that either uncertainty or uncertainty-diversity greatly outperforms random selection. Among them, uncertainty-diversity sampling always performs best,

which verifies the effectiveness of considering the diversity in uncertainty sampling. The success of the diversity measurement is mainly due to the fact that it can effectively avoid duplicate annotation. For example, while the word "企業/enterprise" occurs 392 times in the newly-obtained training data of CityU after using uncertainty sampling, it only occurs 144 times after using uncertainty-diversity sampling.

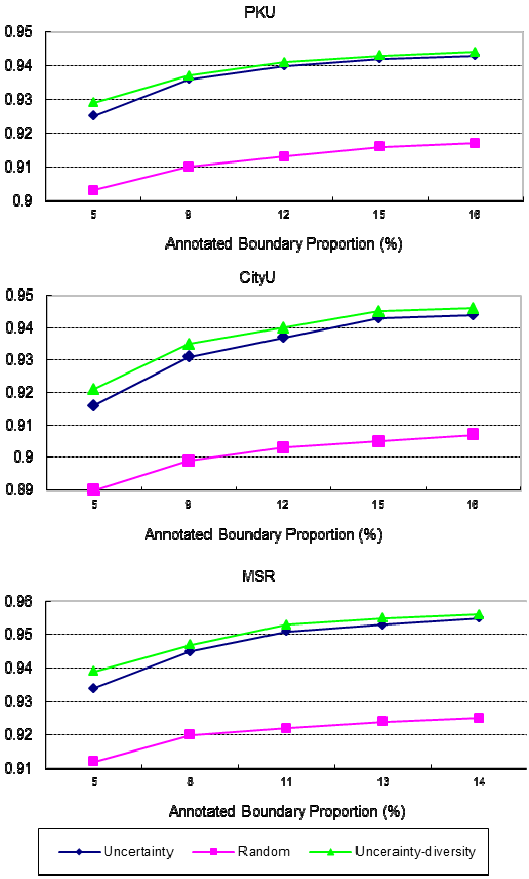

Figure 2: Performance (F1-score) comparison of active learning with different sampling strategies

WBA on Annotation Effort

In this experiment, we randomly draw three different data sets from training data in PKU and ask three students to annotate. Here, each data set has 50 sentences, containing 2186, 2556 and 2528 characters respectively. For a quick annotation, we design an annotation tool where the boundary

between two neighbouring Chinese characters is shown for annotation as a word boundary or not. In particular, three different strategies are used to annotate the data: the first one annotates all sentences; the second one annotates the sentences that contain one or more uncertain boundaries, and the third one only annotates uncertain boundaries (our WBA model).

Here, the main differences between the second and third ones are the context range of the uncertain boundaries. The second one needs the whole sentence as its context and needs to annotate the whole sentence. The third one (used in our approach) only needs part of the sentence as the context (see Section 3.2 in detail) and thus only needs to annotate the uncertain boundary. Table 1 shows real annotation time and the proportion to that of annotating all sentences. From this table, we can see that our active learning approach could save averagely 85% of annotation time and is obviously preferable to the way of annotating the whole sentence.

	All Sentences		Selected Sentences		Selected Boundaries (Our approach)	
	Time	Proportion	Time	Proportion	Time	Proportion
Data Set 1	1232s	100%	790s	64.1%	239s	19.4%
Data Set 2	1746s	100%	1162s	66.6%	320s	18.3%
Data Set 3	1967s	100%	1124s	57.1%	178s	9.0%
AVERAGE	*1648s*	*100%*	*1025s*	*62.6%*	*246s*	*15.6%*

Table 1: Time of annotating three different data sets using different strategies. **All Sentences**: annotating all sentences in the each data set; **Selected Sentences**: annotating only the sentences containing uncertain boundaries; **Selected Boundaries**: annotating only the uncertain boundaries.

5 Conclusion

To our best knowledge, this is the first work in successfully employing active learning on Chinese word segmentation. In particular, our active learning approach aims to annotate only uncertain boundaries with the context automatically labelled. This is achieved via a WBA (Word Boundary Annotation) model. Besides, an efficient diversity measurement is proposed to further reduce the annotation effort. Experimental results on the SIGHAN Bakeoff 2 dataset demonstrate that our active learning approach can greatly reduce the annotation effort with little loss in performance.

Compared to existing studies on active learning for Chinese word segmentation, our approach is unique in two aspects: annotating only the uncertain boundaries instead of the whole sentence, and the diversity measurement, both of which have shown to fairly reduce the annotation cost.

Acknowledgments

The research work described in this paper has been partially supported by three NSFC grants, No.61003155, No.60873150 and No. 61273320, one National High-tech Research and Development Program of China No.2012AA011102, Open Projects Program of National Laboratory of Pattern Recognition. We also thank the three anonymous reviewers for their helpful comments.

References

Andrew G. 2006. A Hybrid Markov/Semi-Markov Conditional Random Field for Sequence Segmentation. In Proceedings of EMNLP-2006, pages 465–472.

Chan Y. and H. Ng. 2007. Domain Adaptation with Active Learning for Word Sense Disambiguation. In Proceedings of ACL-2007, pages 49-56.

Chen J., A. Schein, L. Ungar and M. Palmer. 2006. An Empirical Study of the Behavior of Active Learning for Word Sense Disambiguation. In Proceedings of HLT/NAACL-2006, pages 120–127.

Emerson T. 2005. The Second International Chinese Word Segmentation Bakeoff. In Proceedings of SIGHAN-2005, pages 123-133.

Fu G., C. Kit, J.Webster. 2008. Chinese Word Segmentation as Morpheme-based Lexical Chunking. Information Sciences, 178,(9): 2282-2296.

Fujii A., K. Inui, T. Tokunaga and H. Tanaka. 1998. Selective Sampling for Example-based Word Sense Disambiguation. Computational Linguistics, 24(4): 573-597.

Huang C. and N. Xue. 2012. Words Without Boundaries: Computational approaches to Chinese Word Segmentation. Language and Linguistics Compass. (to appear).

Huang C., P. Šimon, S. Hsieh and L. Prevot. 2007. Rethinking Chinese Word Segmentation: Tokenization, Character Classification, or Wordbreak identification. In Proceedings of ACL-2007 (poster), pages 69-72.

Jiang W., L. Huang, Q. Liu and Y. Lv. 2008. A Cascaded Linear Model for Joint Chinese Word Segmentation and Part-of-Speech Tagging. In Proceedings of ACL-HLT-2008, pages 897–904.

Kudo T. 2005. CRF++: http://crfpp.sourceforge.net/

Laws F. and H. Schütze. 2008. Stopping Criteria for Active Learning of Named Entity Recognition. In Proceedings of COLING-2008, pages 465-472.

Lewis D. and W. Gale. 1994. A Sequential Algorithm for Training Text Classifiers. In Proceedings of SIGIR-1994, pages 3-12.

Li S. and C. Huang. 2009. Word Boundary Decision with CRF for Chinese Word Segmentation. In Proceedings of PACLIC-2009, pages 726-732.

Li S., S. Ju, G. Zhou, and X. Li. 2012. Active Learning for Imbalanced Sentiment Classification. In Proceedings of EMNLP-CoNLL-2012, pages 139-148.

Liere R. and P. Tadepalli. 1997. Active Learning with Committees for Text Categorization. In Proceedings of AAAI-1997, pages 591-596.

McCallum A. and Nigam K. 1998. Employing EM in Pool-based Active Learning for Text Classification. In Proceedings of ICML-1998, pages 350-358.

Ng H. and J. Low. 2004. Chinese Part-of-speech Tagging: One-at-a-time or All-at-once? Word-Based or Character-based. In Proceedings of EMNLP-2004, pages 277–284.

Sassano M. 2002. An Empirical Study of Active Learning with Support Vector Machines for Japanese Word Segmentation. In Proceedings of ACL-2002, pages 505-512.

Settles B. and M. Craven. 2008. An Analysis of Active Learning Strategies for Sequence Labeling Tasks. In Proceedings of EMNLP-2008, pages 1070–1079.

Seung H., M. Opper and H. Sompolinsky. 1992. Query by Committee. In Proceedings of the ACM Workshop on Computational Learning Theory, pages 287–294.

Shen D., J. Zhang, J. Su, G. Zhou and C. Tan. 2004. Multi-criteria-based Active Learning for Named Entity Recognition. In Proceedings of ACL-2004, pages 589-596.

Sproat R. and C. Shih. 1990. A Statistical Method for Finding Word Boundaries in Chinese Text. Computer Processing of Chinese and Oriental Languages, 4(4):336–351.

Sun M., D. Shen and B. Tsou. 1998. Chinese Word Segmentation without Using Lexicon and Handcrafted Training Data. In Proceedings of ACL-COLING-1998, pages 1265-1271.

Tang B., X. Wang and X. Wang. 2009. Chinese Word Segmentation Based on Large Margin Methods. International Journal of Asian Language Processing, 19(1): 55-68.

Tseng H., P. Chang, G. Andrew, D. Jurafsky and C. Manning. 2005. A Conditional Random Field Word Segmenter for Sighan Bakeoff 2005. In Proceedings of SIGHAN-2005, pages 168-171.

Wang K., C. Zong, and K. Su. 2010. A Character-Based Joint Model for Chinese Word Segmentation. In Proceedings of COLING-2010, pages 1173-1181.

Wang K., C. Zong and K. Su. 2012. Integrating Generative and Discriminative Character-Based Models for Chinese Word Segmentation. ACM Transactions on Asian Language Information Processing, vol.11, no.2, pages 7:1-7:41.

Wang Z., C. Huang and J. Zhu. 2008. The Character-based CRF Segmenter of MSRA&NEU for the 4th Bakeoff. In Proceedings of SIGHAN-2008, pages 98-108.

Xue N. 2003. Chinese Word Segmentation as Character Tagging. Computational Linguistics and Chinese Language Processing, 8 (1): 29-48.

Zhang H., H. Yu, D. Xiong and Q. Liu. 2003. HHMM-based Chinese Lexical Analyzer ICTCLAS. In Proceedings of SIGHAN-2003, pages 184-187.

Zhang R., G. Kikui and E. Sumita. 2006. Subword-Based Tagging for Confidence-Dependent Chinese Word Segmentation. In Proceedings of ACL-COLING-2006, pages 961-968.

Zhang Y. and S. Clark. 2007. Chinese Segmentation with a Word-based Perceptron Algorithm. In Proceedings of ACL-2007, pages 840-847.

Zhao H. and C. Kit. 2008. Unsupervised Segmentation Helps Supervised Learning of Character Tagging for Word Segmentation and Named Entity Recognition. In Proceedings of SIGHAN-2008, pages 106-111.

Zhao H. and C. Kit. 2011. Integrating Unsupervised and Supervised Word Segmentation: The Role of Goodness Measures. Information Sciences, 181(1):163-183.

Zhao H., C. Huang, M. Li and B. Lu. 2006. Effective Tag Set Selection in Chinese Word Segmentation via Conditional Random Field Modeling. In Proceedings of PACLIC-2006, pages 87-94.