CRF-based Experiments for Cross-Domain Chinese Word Segmentation at CIPS-SIGHAN-2010

Xiao Qin, Liang Zong, Yuqian Wu, Xiaojun Wan and Jianwu Yang
Institute of Computer Science and Technology
Peking University, China, 100871
{qinxiao,zongliang,wuyuqian,wanxiaojun,yangjianwu}
@cist.pku.edu.cn

Abstract

This paper describes our experiments on the cross-domain Chinese word segmentation task at the first CIPS-SIGHAN Joint Conference on Chinese Language Processing. Our system is based on the Conditional Random Fields (CRFs) model. Considering the particular properties of the out-of-domain data, we propose some novel steps to get some improvements for the special task.

1 Introduction

Chinese word segmentation is one of the most important tasks in the field of Chinese information processing and it is meaningful to intelligent information processing technologies. After a lot of researches, Chinese word segmentation has achieved a high accuracy. Many methods have been presented, among which the CRFs model has attracted more and more attention. Zhao's group used the CRFs model in the task of Chinese word segmentation in Bakeoff-4 and they ranked at the top in all closed tests of word segmentation (Zhao and Kit, 2008). The CRFs model has been widely used because of its excellent performance. However, finding a better segmentation algorithm for the out-of-domain text is the focus of CIP-SIGHAN-2010 bakeoff.

We still consider word segmentation as a sequence labeling problem. What we concern is how to use the unlabeled corpora to enrich the supervised CRFs learning. So we take some strategies to make use of the information of the texts in the unlabeled corpora.

2 System Description

In this section, we will describe our system in details. The system is based on the CRFs model and we propose some novel steps for some improvements. It mainly consists of three steps: preprocessing, CRF-based labeling, and re-labeling.

2.1 Preprocessing

This step mainly includes two operations. First, we should cut the whole text into a series of sentences. We regard '。', '？', '！' and '；' as the symbols of the boundary between sentences. Then we do atomic segmentation to all the sentences. Here Atomic segmentation represents that we should regard the continuous non-Chinese characters as a whole. Take the word 'computer' as an example, we should regard 'computer' as a whole, but not treat it as 8 separate letters of 'c', 'o', 'm', 'p', 'u', 't', 'e', and 'r'.

2.2 CRF-based Labeling

Conditional random field (CRF) is an extension of both Maximum Entropy Model (MEMs) and Hidden Markov Models (HMMs), which was firstly introduced by Lafferty (Lafferty et al., 2001). It is an undirected graphical model trained to maximize the conditional probability of the desired outputs given the corresponding inputs. This model has achieved great successes in word segmentation.

In the CRFs model, the conditional distribution $P(y|x)$ of the labels Y givens observations X directly is defined:

$$P(y/x) = \frac{1}{Z_x}\exp\{\sum_{t=1}^{T}\sum_{k}\lambda_k f_k(y_{t-1}, y_t, x, t)\}$$

y is the label sequence, x is observation sequence, Z_x is a normalization term that makes the probability of all state sequences sum to one; $f_k(y_{t-1}, y_t, t)$ is often a binary-valued feature function and λ_k is the weight of f_k.

In our system, we choose six types of tags according to character position in a word. According to Zhao's work (Zhao et al., 2006a), the 6-tag set enables our system to generate a better CRF model than the 4-tag set. In our experiments, we test both the 6-tag set and the 4-tag set, and the 6-tag set truly has a better result. The 6-tag set is defined as below:

T = {B, B2, B3, M, E, S}

Here B, B2, B3, M, E represent the first, second, third, continuing and end character positions in a multi-character word, and S is the single-character word tag.

We adopt 6 n-gram feature templates as features. Some researches have proved that the combination of 6-tag set and 6 n-gram feature template can achieve a better performance (Zhao et al., 2006a; Zhao et al., 2006b; Zhao and Kit, 2007).

The 6 n-gram feature templates used in our system are C_{-1}, C_0, C_1, $C_{-1}C_0$, C_0C_1, $C_{-1}C_1$. Here C stands for a character and the subscripts -1, 0 and 1 stand for the previous, current and next character, respectively.

Furthermore, we try to take advantage of the types for the characters. For example, in our system D stands for the date, N stands for the number, L stands for the letter, P stands for the punctuation and C stands for the other characters. Introducing these features is beneficial to the CRFs learning.

2.3 Re-labeling step

Since the unlabeled corpora belong to different domains, traditional methods have some limitations. In this section, we propose an additional step to make good use of the unlabelled data for this special task. This step is based on the outputs of the CRFs model in the previous step.

After CRFs learning, we get a training model. With this model, we can label the literature, computer, medicine and finance corpora. According to the outputs of the CRFs model, we choose some labeled sentences with high confidence and add them to the training corpus. Here the selection of high confidence must guarantee that the probability of the sentences selected being correct segmentations is rather high and the number of the sentences selected is not too little or they will make no difference to the generation of the new CRF model. Since the existing training model does not contain the information in the out-of-domain data, we treat the labeled sentences with high confidence as additional training corpus. Then we re-train the CRFs model with the new training data. With the training data extracted from different domains, the training model incorporates more cross-domain information and it can work better in the corresponding cross-domain prediction task.

3 Experiments

3.1 Experiment Setup

There are two sources for the corpora: the training corpora and the test corpus. And in the training corpora, there exist two types of corpus in this task. The labeled corpus is Chinese text which has been segmented into words while the unlabelled corpus covers two domains: literature and computer science. The test corpus contains 4 domains, which are literature, computer science, medicine and finance.

There are four evaluation metrics used in this bake-off task: Precision, Recall, F1 measure (F1 = 2RP/(R+P)) and OOV measure, where R and P are the recall and precision of the segmentation and OOV (Out-Of-Vocabulary Word) is a word which occurs in the reference corpus but does not occur in the labeled training corpus.

Our system uses the CRF++ package Version 0.49 implemented by Taku Kudo [1] from sourceforge.

3.2 Results and Discussions

We test the techniques described in section 2 with the given data. Now we will show the results of each operation.

3.2.1 Preprocessing

As we have mentioned in section 2.1, the first step is to cut the text into a series of sentences.

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

Then we should give each character in one sentence a label. Before this step, it is necessary to do atomic segmentation. And we will regard the continuous non-Chinese characters as a whole and give the whole part a single label. This is meaningful to those corpora containing a lot of English words. Due to the diversity of the English words, segmenting the sentences with a lot non-Chinese characters correctly is rather difficult only through CRF learning. We should do atomic segmentation to all training and test corpora. This may achieve a higher accuracy in a certain degree.

The results of word segmentation are reported in Table 1. 'Clouse+/-' indicates whether text clause has been done.

Table 1: Results with clause and without clause

	corpus	Precision	Recall	F
Literature	Clause+	0.922	0.916	0.919
	Clause-	0.921	0.915	0.918
Computer	Clause+	0.934	0.939	0.937
	Clause-	0.934	0.939	0.936
Medicine	Clause+	0.911	0.917	0.914
	Clause-	0.509	0.511	0.510
Finance	Clause+	0.940	0.943	0.941
	Clause-	0.933	0.940	0.937

From Table 1, we can see there is some improvement in different degree and the effect in the medicine corpus is the most obvious. So we can conclude that our preprocessing is useful to the word segmentation.

3.2.2 CRF-based labeling

After preprocessing, we can use CRF++ package to learn and test.

The selection of feature template is also an important factor. For the purpose of comparison, we test two kinds of feature templates in our system. The one is showed in Table 2 and the other one is showed in Table 3.

Table 2: Template 1

Unigram
U00:%x[-1,0]
U01:%x[0,0]
U02:%x[1,0]
U03:%x[-1,0]/%x[0,0]
U04:%x[0,0]/%x[1,0]
U05:%x[-1,0]/%x[1,0]
Bigram
B

Table 3: Template 2

Unigram
U00:%x[-1,0]
U01:%x[0,0]
U02:%x[1,0]
U03:%x[-1,0]/%x[0,0]
U04:%x[0,0]/%x[1,0]
U05:%x[-1,0]/%x[1,0]
U10:%x[-1,1]
U11:%x[0,1]
U12:%x[1,1]
U13:%x[-1,1]/%x[0,1]
U14:%x[0,1]/%x[1,1]
U15:%x[-1,1]/%x[1,1]
Bigram
B

Now we will explain the meanings of the templates. Here is an example. In table 4, we show the format of the input file. The first column represents the word itself and the second represents the feature of the word, where there are five kinds of features: date (D), number (N), letter (L), punctuation (P) and others (C). The meanings of the templates are showed in table 5.

Table 4: the format of the input file for CRF

新	C
年	D
讲	C
话	C
（	P
附	C
图	C
片	C
1	N
张	C
）	P

Table 5: the example of the templates

template	Expanded feature
%x[0,0]	图
%x[0,1]	C
%x[1,0]	片
%x[-1,0]	附
%x[-1,0]/ %x[0,0]	附/图
%x[0,0]/ %x[0,1]	图/C

With two different feature templates, we continue our experiments in the four different domains. The segmentation performances of our system on test corpora using different feature templates are presented in Table 6.

Table 6: Results with different feature templates

	corpus	Precision	Recall	F
Literature	T1	0.917	0.909	0.913
	T2	0.922	0.916	0.919
Computer	T1	0.914	0.902	0.908
	T2	0.934	0.939	0.937
Medicine	T1	0.906	0.905	0.905
	T2	0.911	0.917	0.914
Finance	T1	0.937	0.925	0.931
	T2	0.940	0.943	0.941

Here T1 stands for Template 1 while T2 stands for Template 2.

From the Table 4 we can see the second feature templates make the results of the segmentation improved more significantly.

At the same time we need get the outputs with confidence measure by setting some parameters in CRF test.

3.2.3 Re-labeling

As for the outputs with confidence measure generated by previous step, we should do some special processes. Here we set a particular value as our standard and choose the sentences with confidence above the value. As we know, the test corpora are limited, the higher confidence may cause the corpora meeting our requirements are less. The lower confidence may not guarantee the reliability. So the setting of the confidence value is very significant. In our experiments, we set the parameter at 0.8.

Then we add the sentences whose confidence is above 0.8 to the training corpus. We should re-learn with new corpora, generate the new model and re-test the corpora related with 4 domains. The segmentation performances after re-labeling are represented in Table 7.

Table 7: Results with re-labeling and without re-labeling

	corpus	Precision	Recall	F
Literature	Re +	0.922	0.916	0.919
	Re -	0.921	0.916	0.918
Computer	Re +	0.934	0.939	0.937
	Re -	0.932	0.934	0.933
Medicine	Re +	0.911	0.917	0.914
	Re -	0.912	0.918	0.915
Finance	Re +	0.940	0.943	0.941
	Re -	0.937	0.941	0.939

Here Re+/- indicates whether the re-labeling step is to be done.

From the results we know, even though the re-labeling step makes the results in the medicine corpus a little worse, it has much better effect in the other corpora. Overall, the operation of re-labeling is necessary.

3.3 Our results in this bakeoff

In this task, our results are showed in Table 8.

Table 8: our results in this bakeoff

	Precision	Recall	F
Literature	0.922	0.916	0.919
Computer	0.934	0.939	0.937
Medicine	0.911	0.917	0.914
Finance	0.940	0.943	0.941

From Table 6, we can see our system can achieve a high precision, especially in the domains of computer and finance. This proves our methods are fairly effective.

4 Discussion

4.1 Segmentation Features

In our system, we only take advantage of the features of the words. We try to add other fea-

tures to our experiments such as AV feature (Feng et al., 2004a; Feng et al., 2004b; Hai Zhao et al., 2007) with the expectation of improving the results. But the results are not satisfying. We believe that the feature of words frequency may be an important factor, but how to use it is worth studying. So finding some meaningful and effective features is the crucial point.

4.2 OOV

In our system, we do not process the words out of vocabulary in the special way. The recognition of OOV is still a problem. In a word, there is still much to be done to improve our system. In the present work, we make use of some surface features, and further study should be continued to find more effective features.

5 Conclusion

In this paper, we have briefly described the Chinese word segmentation for out-of-domain texts. The CRFs model is implemented. In order to make the best use of the test corpora, some special strategies are introduced. Further improvement is made with these strategies. However, there is still much to do to achieve more improvement. From the results, we got good experience and knew the weaknesses of our system. These all help to improve the performance of our system in the future.

Acknowledgements

The research described in this paper was supported by NSFC (Grant No. 60875033).

References

Hai Zhao, Changning Huang, and Mu Li. 2006. An improved Chinese Word Segmentation System with Conditional Random Field. *Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing*, Sydney, Australia.

Hai Zhao and Chunyu Kit. 2007. Incorporating global information into supervised for Chinese word segmentation. In *PACALING-2007*, Melbourne, Australia.

Hai Zhao, Changning Huang, Mu Li and Bao-Liang Lu. 2006b. Effective tag set selection in Chinese word segmentation via conditional random field modeling. In *PACLIC-20*, Wuhan, China.

John Lafferty, Andrew McCallum, and Fernando Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceeding of ICML 2001*, Morgan Kaufmann, San Francisco, CA

Hai Zhao and Chunyu Kit. 2008. Unsupervised Segmentation Helps Supervised Learning of Character Tagging for Word Segmentation and Named Entity Recognition. *Processing of the Sixth SIGHAN Workshop on Chinese Language Processing*, Hyderabad, India.

Haodi Feng, Kang Chen, Xiaotie Deng, and Weimin Zheng. 2004a. Accessor variety criteria for Chinese word extraction. *Computational Linguistics*.

Haodi Feng, Kang Chen, Chunyu Kit, and Xiaotie Deng. 2004b. Unsupervised segmentation of Chinese corpus using accessor variety. *In First International Joint Conference on Natural Language Processing*. Sanya, Hainan Island, China.