Chinese Word Segmentation based on Mixing Multiple Preprocessor and CRF

Jianping Shen, XuanWang, Hainan Zhao,Wenxiao Zhang
Computer Application Research Center, Shenzhen Graduate School
Harbin Institute of Technology Shenzhen, China, 518055
Email: {jpshen, wangxuan,hnzhao@cs.hitsz.edu.cn} Email: { xiaohit@126.com}

Abstract

This paper describes the Chinese Word Segmenter for our participation in CIPS-SIGHAN-2010 bake-off task of Chinese word segmentation. We formalize the tasks as sequence tagging problems, and implemented them using conditional random fields (CRFs) model. The system contains two modules: multiple preprocessor and basic segmenter. The basic segmenter is designed as a problem of character-based tagging, and using named entity recognition and chunk recognition based on boundary to preprocess. We participated in the open training on Simplified Chinese Text and Traditional Chinese Text, and our system achieved one Rank#5 and four Rank#2 best in all four domain corpus.

1 Introduction

Word is a logical semantic and syntactic unit in natural language (Zhenxing Wang, 2008). Chinese word segmentation is very important for Chinese language processing, which aims to recognize the implicit word boundaries in Chinese text. It is the foundation of most Chinese NLP tasks. In past decades, great success has been achieved in Chinese word segmentation (Nie, et al, 1995; Wang et al, 2000;Zhang, et al, 2002). But there still exist many problems, such as cross-domain performance of Chinese word segmentation algorithms. As the development of the internet, more and more new word has been appearing, Improving the performance of Chinese word segmentation algorithms on OOV (Out-Of-Vocabulary Word, is a word which occurs in the reference corpus but does not occur in the labeled training corpus) is the important research direction. Our system participated in the CIPS-SIGHAN-2010 bake-off task of Chinese word segmentation. And we have done work in dealing with two main sub-tasks: (1) Word Segmentation for Simplified Chinese Text, (2) Word Segmentation for Traditional Chinese Text. Our system formalizes these tasks as consecutive sequence tagging problems, and learns the segmentation using conditional random fields approach. Our system contains two modules, a multiple preprocessor and a basic segmenter. The multiple preprocessor first finds chunks based on boundary dictionary and then uses named entity recognition technology to extract the person, location, organization and special time. The basic segmenter using CRF model is trained to segment the sentence to word which contains one or more characters. The basic segmenter follows the study of Zhenxing Wang, Changning Huang and Jingbo Zhu (2008), but applies more refined features and tags.

The reminder of the paper is organized as follows. In section 2, we briefly describe the task and the details of our system. The experimental results are discussed in section 3. In section 4 we put forward our conclusion.

2 System Description

In this section we describe our system in more detail. The Figure1 is the frame of our system. It contains two modules: multiple preprocessor and basic segmenter.

2.1 Multiple Preprocessor

The preprocessor contain two modules: chunking based on boundary dictionary and NE Reorganization.

2.1.1 Chunking

In one sentence, there are always some characters or words, such as “是”, ”的”,”与”,”基于”, the character adjacent them can not together with them. We define these characters or words as boundary word. We built a boundary

dictionary manual, which contains about 100 words. Once a

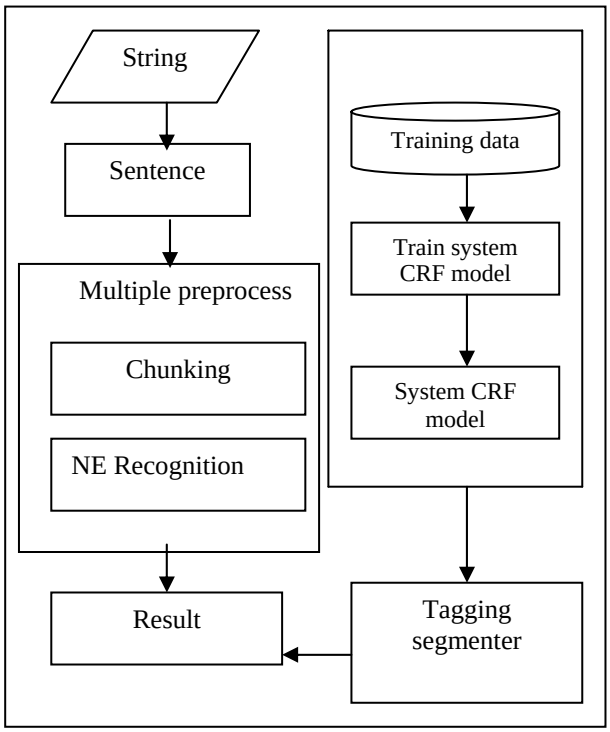

Figure1. Chinese Word Segmenter

sentence input, our system finds boundary words in the sentence first. For example, such as, "欧元区和欧盟成员国的财政部长分别于 15 日和 16 日在布鲁塞尔召开月度例会". In this sentence we can find the boundary word"和" "的" "于" "在" " 分别". Then chunking result is shown below in Figure 2. Using "[ ]" to mark up the chunks.

[欧元区] 和 [欧盟成员国] 的 [财政部长] 分别 于 [15 日] 和 [16 日] 在 [布鲁塞尔] [召开月度例会]

Figure 2. a sentence with chunk in data set

Chunking is very useful to improve the precision of segmentation. Especially when lacking enough training corpus for training CRF model. It can improve the out-of-vocabulary (OOV) word Recall and Precision on cross-domain Chinese word segmentation.

2.1.2 NE Recognition

We will recognize the named entities such as persons, locations organizations. We perform a process of the named entities recognition with forward-backward maximum matching algorithm based on entity dictionary. The dictionary contain location dictionary, person dictionary, family name dictionary, organization dictionary, country dictionary (Jianping Shen and Xuan Wang, 2010). For example, a sentence, "东海证券分析师王万金表示，该结果说明中国南车的价值已被市场所认可". And the processor will find out the location"东海","中国", the family name "王" and person "万金". The NE recognition result is shown below in Figure 3.

[东海]/loc 证券分析师 [王]/fam [万金]/per 表示，该结果说明 [中国]/ [南车]/org 的价值已被市场所认可.

Figure 3.sentence with NE recognition in data set

The location tag with "[ ]/loc", family name tag with "[ ]/fam", person tag with "[ ]/per", organization tag with "[ ]/org".

2.2 Basic Segmenter

We model the segment task as the consecutive sequence labeling problems, such as chunking, and named entity recognition, and train the basic segmenter using conditional random fields approach (Lafferty et al., 2001).

2.2.1 Conditional Random Fields

CRF models are conditional probabilistic sequence and undirected graphical models

CRF models hold two natures. First is the conditional nature, second the exponential nature. The conditional nature of the distribution over label sequence allows CRF models to model real-world data in which the conditional probability of a label sequence can depend on non-independent and interacting features of the observation sequence. The exponential nature of the distribution enables features of different states to be traded off against each other, weighting some states in a sequence as being more important than other states. Following Lafferty et al. and Hanna Wallach, the exponential distribution chosen by John Lafferty et al. is shown as follow:

$$p_\theta(y \mid x) \propto \exp\Big(\sum_{e \in E,k} \lambda_k f_k(e, y|_e, x) + \sum_{v \in V,k} \mu_k g_k(v, y|_v, x)\Big)$$

$$= \exp\Big(\sum_i \sum_k \lambda_k f_k(y_{i-1}, y_i, x)$$

$$+\sum_i \sum_k \mu_k g_k(y_{i,} x)) \quad (1)$$

Where

$$f_{y',y}(y_u, y_v, x) = \begin{cases} 1 & if\ y_u = y'\ and\ y_v = y \\ 0 & otherwise \end{cases}$$

And

$$g_{y,x}(y_v, x) = \begin{cases} 1 & if\ y_v = y\ and\ x_v = x \\ 0 & otherwirse \end{cases}$$

In this situation, the parameters $\lambda_{y',y}$ and $\mu_{y,x}$ corresponding to these features are equivalent to the logarithms of the HMM transition and emission probabilities $p(y'|y)$ and $p(x|y)$. The parameter of the model can be estimated in many ways, such as GIS, IIS, L-BFGS etc.

2.2.2 Segment base on CRF model

When a sentence or chunk (which get from the preprocessor) input, it will be split to the sequences shown in Figure 4.

chunk	sequence
欧盟成员国	欧 盟 成 员 国

Figure 4. Chunk and sequence

Every character in input sentences will be given a label which indicates whether this character is a word boundary. Our basic segmenter is almost the same as the system described in (Zhao et al., 2006) which is learned from training corpus. The CRF model we use is implemented with CRF++ 0.51. The parameters of the CRF segmenter are set as defaults.

Under the CRF tagging scheme, each character in one sentence will be given a label by CRF model to indicate which position this character occupies in a word. In our system, CRF tag set is proposed to distinguish different positions in the multi-character words when the word length is less than 6, namely 6-tag set {B, B2, B3, M, E, O}(Zhenxing Wang ,2008).We defined that B stands for the first and E stands for the last position in a multi-character word. S stands up a single-character word. B2 and B3 stand for the second and the third position in a multi-character word. M stands for the fourth or more rear position in a multi-character word, whose length is larger than four-character. Then we add the entity tag set {B-entity, I-entity, E-entity}. B-entity stands for the first character in a named entity, E-entity stands for the last character in a named entity, and I-entity stands for the other character in a named entity.

We use a greedy forward procedure to select a better feature sets for the segementer according to the evaluation results in the development set. We first start from a basic feature set, and then add each feature outside the basic set and remove each feature inside the basic set one by one to check the effectiveness of each feature by the performance change in the development set. This procedure is repeated until no feature is added or removed or the performance is not improved. The selected features are listed below:

- C_n (n=-2,-1, 0, 1, 2)
- C_nC_{n+1} (n=-1,0)
- $C_{n-1}C_nC_{n+1}$ (n=-1,0,1)
- $C_{n-2}C_{n-1}C_nC_{n+1}$ (n=0,1)

Where C refer to the tag of each character, and C_0 denotes current character and Cn(C-n) denotes the character n positions to the right (left) of current character.

2.2.3 Post-processing

We can obtain the preliminary results through the CRF model-based Segment, but there are some missed or incorrect cases for the digit, English word. For example “the sighan” maybe segment to “th e sig han”, so we will re-segment the “th e sig han” as “the sighan”.

3 Performance and Analysis

In this section we will present our experimental results for these two subtasks. For the Word Segmentation for Simplified Chinese Text subtask, comparing the performance of these four domains, we find that the performance of computer and finance are better than literature and medical. We can find that the OOV RR of literature and medical are lower than the computer and finance. In the test data set, there are many Out-of-vocabulary(OOV), especially the disease. In medical domain, there are many diseases which do not appear in the corpus, and there is the proper name. The segment often can’t recognize disease well, so we add a post-processing procedure, using domain dictionary for medicine, is used to increase the recall

measure. The result for medical is shown in Table 2.

domain	R	P	F1	OOV RR	IV RR
literature	0.836	0.841	0.838	0.609	0.853
computer	0.951	0.951	0.932	0.77	0.983
medical	0.839	0.832	0.836	0.796	0.866
finance	0.893	0.896	0.894	0.796	0.902

Table 1: Performance of the four domain Simplified Chinese test data set

R	P	F1	OOV RR	IV RR
0.894	0.882	0.888	0.683	0.901

Table 2: Performance of medical test data set with post-processing using domain dictionary

Word Segmentation for Traditional Chinese Text subtask. We use a Traditional and Simplified Dictionary to translate the named entity dictionary, boundary dictionary from Simplified to Traditional. And then we use our system to segment the traditional test data set. The results are shown in Table 3.

domain	R	P	F1	OOV RR	IV RR
literature	0.868	0.802	0.834	0.503	0.905
computer	0.875	0.829	0.851	0.594	0.904
medical	0.879	0.814	0.846	0.480	0.912
finance	0.832	0.760	0.794	0.356	0.866

Table 3: Performance of four domain Traditional Chinese test data set

4 Conclusion

Through the CIPS-SIGHAN bakeoff, we find our system is effective. And at the same time, we also find some problems of us. Our system still can't performance very good in cross-domain. Especially the Out-of-vocabulary (OOV) recognition. From the experiment we can see that using domain dictionary is a good idea. In the future we will do more work in post-processing. The bakeoff points out the direction for us to improve our system.

References

Huipeng Zhang, Ting Liu, Jinshan Ma, Xiantao Liao,Chinese *Word Segmentation with Multiple Postprocessors in HIT-IRLab, Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing*, Jeju Island, Republic of Korea October 11-13, 2005

Hai Zhao, Changning Huang et al. 2006. *Effective Tag Set Selection in Chinese Word Segmentation via Conditional Random Field Modeling. In Proceedings of PACLIC-20*. pages 87-94. Wuhan, China, Novemeber.

Zhenxing Wang; Changning Huang; Jingbo Zhu. *The Character-based CRF Segmenter of MSRA&NEU for the 4th Bakeoff. The Sixth SIGHAN Workshop for Chinese Language was be held in conjunction with IJCNLP 2008*, in Hyderabad, India, January 11-12, 2008.

Wei Jiang Jian Zhao Yi Guan Zhiming Xu. *Chinese Word Segmentation based on Mixing Model. Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing,* Jeju Island, Republic of Korea October 11-13, 2005

Nie, Jian-Yuan, M.-L. Hannan and W.-Y. Jin. 1995.*Unknown word detection and segmentation ofChinese using statistical and heuristic knowledge.Communication of COLIPS*, 5(1&2): 47-57.

Wang, Xiaolong, Fu Guohong, Danial S.Yeung, James N.K.Liu, and Robert Luk. 2000. *Models and algorithms of Chinese word segmentation. In: Proceedings of the International Conference on Artificial Intelligence (IC-AI'2000)*, Las Vegas, Nevada, USA, 1279-1284.

Lafferty, J. and McCallum, A. and Pereira, F. *Conditional random fields: Probabilistic models for segmenting and labeling sequence data. MACHINE LEARNING-INTERNATIONAL WORKSHOP THEN CONFERENCE*. 2001, 282-289

Hanna Wallach, *Efficient Training of Conditional Random Fields, In Proceedings of the 6th Annual CLUK Research Colloquium*, 2002.