A Hybrid Approach to Chinese Word Segmentation around CRFs

ZHOU Jun-sheng[1, 2] **DAI Xin-yu**[1] **NI Rui-yu**[1] **CHEN Jia-jun**[1]

[1]Department of Computer Science and Technology, Nanjing University, Nanjing, 210093 CHINA
[2]Deptartment of Computer Science, Nanjing Normal University, Nanjing, 210097 CHINA
{Zhoujs, dxy, niry, chenjj}@nlp.nju.edu.cn

Abstract

In this paper, we present a Chinese word segmentation system which is consisted of four components, i.e. basic segmentation, named entity recognition, error-driven learner and new word detector. The basic segmentation and named entity recognition, implemented based on conditional random fields, are used to generate initial segmentation results. The other two components are used to refine the results. Our system participated in the tests on open and closed tracks of Beijing University (PKU) and Microsoft Research (MSR). The actual evaluation results show that our system performs very well in MSR open track, MSR closed track and PKU open track.

1 Introduction

Word segmentation is the first step in Chinese NLP, but segmentation of the Chinese text into words is a nontrivial task. Three difficult tasks, i.e. ambiguities resolution, named entity recognition and new word identification, are the key problems to word segmentation in Chinese.

In this paper, we report a Chinese word segmentation system using a hybrid strategy. In our system, texts are segmented in four steps: basic segmentation, named entity recognition, error-driven learning and new word detection. The implementations of basic segmentation component and named entity recognition component are both based on conditional random fields (CRFs) (Lafferty et al., 2001), while the Error-Driven learning component and new word detection component use statistical and rule methods. We will describe each of these steps in more details below.

2 System Description

2.1 Basic segmentation

We implemented the basic segmentation component with linear chain structure CRFs. CRFs are undirected graphical models that encode a conditional probability distribution using a given set of features. In the special case in which the designated output nodes of the graphical model are linked by edges in a linear chain, CRFs make a first-order Markov independence assumption among output nodes, and thus correspond to finite state machines (FSMs). CRFs define the conditional probability of a state sequence given an input sequence as

$$P_{\mathrm{A}}(s \mid o) = \frac{1}{Z_o} \exp\left(\sum_{t=1}^{T} \sum_{k=1}^{K} \lambda_k f_k(s_{t-1}, s_t, o, t) \right)$$

Where $f_k(s_{t-1}, s_t, o, t)$ is an arbitrary feature function over its arguments, and λ_{k} is a learned weight for each feature function.

Based on CRFs model, we cast the segmentation problem as a sequence tagging problem. Different from (Peng et al., 2004), we represent the positions of a *hanzi* (Chinese character) with four different tags: B for a *hanzi* that starts a word, I for a *hanzi* that continues the word, F for a *hanzi* that ends the word, S for a *hanzi* that occurs as a single-character word. The basic segmentation is a process of labeling each *hanzi* with a tag given the features derived from its surrounding context. The features used in our experiment can be broken into two categories: character features and word features. The character features are instantiations of the following templates, similar to those described in (Ng and Jin, 2004), *C* refers to a Chinese *hanzi*.

(a) *Cn (n = −2,−1,0,1,2)*
(b) *CnCn+1(n = −2,−1,0,1)*
(c) *C−1C1*
(d) *Pu(C0)*

In addition to the character features, we came up with another type word context feature which was found very useful in our experiments. The feature captures the relationship between the *hanzi* and the word which contains the *hanzi*. For a two-*hanzi* word, for example, the first *hanzi* "连" within the word "连续" will have the feature WC0=TWO_F set to 1, the second *hanzi* "续" within the same word "连续" will have the feature WC0=TWO_L set to 1. For the three-*hanzi* word, for example, the first *hanzi* "梳" within a word "梳妆镜" will have the feature WC0=TRI_F set to 1, the second *hanzi* "妆" within the same word "梳妆镜" will have the feature WC0=TRI_M set to 1, and the last *hanzi* "镜" within the same word "梳妆镜" will have the feature WC0=TRI_L set to 1. Similarly, the feature can be extended to a four-*hanzi* word.

2.2 Named Entity recognition

After basic segmentation, a great number of named entities in the text, such as personal names, location names and organization names, are not yet segmented and recognized properly. So the second step we take is named entity recognition based on CRFs. In contrast to Chinese personal names and location name, the recognition of Chinese organization names is a difficult task. Especially in Microsoft Research corpus, the whole organization name, such as "国际竹藤组织", "北京航空航天大学国家软件开发环境重点实验室" and so on, is regarded as a single word. In this section, we only present our approach for organization name recognition.

The important factor in applying CRFs model to organization name recognition is how to select the proper features set. The constitution of Chinese organization is very complicated, and most organization names do not have any common structural characteristics except for containing some feature words, such as 公司, 学校 and so on. But as a proper noun, the occurrence of an organization name has the specific context. The context information of organization name mainly includes the boundary words and some title words (e.g. 局长、董事长). By analyzing a large amount of organization name corpus, we find that the indicative intensity of different boundary words vary greatly. So we divide the left and right boundary words into two classes according to the indicative intensity. Accordingly we construct the four boundary words lexicons. To solve the problem of the selection and classification of boundary words, we make use of mutual Information *I(x, y).* If there is a genuine association between x and y, then I(x,y*)* >> 0. If there is no interesting relationship between x and y, then I(x,y)≈0. If x and y are in complementary distribution, then I(x,y) << 0. By using mutual information, we compute the association between boundary word and the type of organization name, then select and classify the boundary words.

In order to increase the precision of organization name recognition, we still introduce the "forbidden word" feature that would prevent some words from being recognized as component of organization name.

For we know that some words, such as “当然”, “即使”, are impossible to occur in organization name, we collected these words to formed a “forbidden words” lexicon. Based on the consideration given in preceding section, we constructed a set of atomic feature patterns, listed in table 2. Additionally, we defined a set of conjunctive feature patterns, which could form effective feature conjunctions to express complicated contextual information.

2.3 Error-driven learning

As a method based on statistics, no matter how well a CRFs model is constructed, some obviously errors always occurred because of the sparseness of training data. For this reson, error-driven learning method (Brill, 1995) is adopted to refine the segmentation result in this bakeoff in three steps:

1) Based on CRFs model, we segment the *training data* which has been removed all the space between words. Based on the comparison of the segmentation result with the original training data, the difference between them will be extracted. If a difference occurs more than one time, an error-driven rule will be constructed. The rule is described as: α → w1w2w3, w1w2w3 is the segmentation of α in training data. We named this rule set constructed by this step *CRF-Ruleset*.

2) Based on FMM&BMM, we segment the *training data* which has been removed all the space between words. As we know, overlapping ambiguity strings can be found through FMM&BMM, and the true segmentation of such *OASs* can be found in training data. If an OAS string has unique segmentation, a rule α → w1w2w3 was constructed. We called the rule set constructed in this step *OAS-Ruleset*.

3) In the testing data, if there is the same string as α in *CRFs-Ruleset* or *OAS-Ruleset*, it will be segmented as w1w2w3 according to the rule α → w1w2w3.

For example, in the PKU testing data, through error-driven learning, we can segment the string “邓小平理论” as “邓小平理论” while this string is always segmented wrong as “邓 小平 理论” segmented by CRFs model.

In other words, error-driven learning can always can be seen as a consistency check. It assures the consistency of the segmentation of the training data and testing data when some strings such as “邓小平理论” occur in both.

2.4 New word detection

CRFs segmentation model can gives good performance on OOV words identification. But there are still some new words that have not been recognized. So an additive new words recognizer is adopted (Chen, 2003).

In-word probability of each character is used for new word detection. The in-word probability of a character is a probability that the character occurs as a part of a word of two or more characters. And the in-word probability of a character is trained from the training data and is calculated as follows:

$$P_{in-word}(C) = \frac{Number\ of\ C\ Occurrence\ in\ words}{Number\ of\ C\ Occurrence}.$$

The consecutive single characters are combined into a new word if the in-word probability of each single character is over a threshold. Obviously, the value of the threshold is the key to the performance of this new words recognizer. Same as (Chen, 2003), we divided the training data as training data and developing data to find an exactly value of the threshold. For this bakeoff, we set the threshold of PKU data as 0.86 and that of MSR data as 0.88. Some new words such as “车购费” “互致” “亚欧” were recognized by this recognizer.

3 Experimental results

We participated in the four GB tracks in the second international Chinese word segmentation bakeoff: PKU-open, PKU-closed, MSR-open, MSR-closed. In the closed tracks, we used the dictionary with the words appearing in the training corpus and didn’t conduct the process of named entity recognition. In the open tracks, we employed a dictionary of 134,458 entries. The size of training data used in the open tracks is same as the closed tracks. Except for a dictionary with more vocabulary, we have not employed any other special resources in the open tracks. Table 1 shows the performance of our system in the bakeoff.

	PKU (open)	PKU (closed)	MSR (open)	MSR (closed)
Precision	0.970	0.950	0.971	0.956
Recall	0.964	0.941	0.959	0.959
F	0.967	0.946	0.965	0.957
OOV	0.058	0.058	0.026	0.026
Recall on OOV	0.864	0.813	0.785	0.496

Table 1: Official Bakeoff Outcome

It’s a pity that we make a careless mistake (a program bug) which led to 752 left quotation marks concatenated to the words following it in the closed and open tracks on Microsoft research corpus. With the problem fixed, the actual results of the official test data are better than any other system, as shown in Table 2.

	MSR (open)	MSR (closed)
Precision	0.978	0.957
Recall	0.976	0.976
F	0.977	0.966
OOV	0.026	0.026
Recall on OOV	0.772	0.387
Recall on In-Voc	0.982	0.992

Table 2 Actual evaluation on MSR corpus

4 Conclusion

Our open and closed GB track experiments show that its performance is competitive. The most important advantage of our system is the ability to cope with the unknown words. Especially in the open track on the Microsoft research corpus, the recall on OOV of our system achieves 77.2%, higher than any other system. In future work, we would attempt to generalize the ideas of large-margin to CRFs model, leading to new optimal training algorithms with stronger guarantees against overfitting.

References

Eric Brill , 1995. Transformation Based Error Driven Learning and Natural Language Processing : A Case Study in Part of Speech Tagging , Computational Linguistics , V21. No. 4 ,

J. Lafferty, A. McCallum, and F. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In Proc. ICML 01.

Aitao Chen. 2003. Chinese Word Segmentation Using Minimal Linguistic Knowledge. In *Proceedings of the Second SIGHAN Workshop on Chinese Language Processing*.

Ng, Hwee Tou and Jin Kiat Low. 2004. Chinese Part-of-Speech Taging: One-at-a-Time or All at Once? Word-based or Character based? In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Spain.

Peng, Fuchun, Fangfang Feng, and Andrew McCallum. 2004. Chinese Segmentation and New Word Detection using Conditional Random Fields . In *Proceedings of the Twentith International Conference on Computaional Linguistics*, pages 562–568.