Term Contributed Boundary Feature using Conditional Random Fields for Chinese Word Segmentation Task

Tian-Jian Jiang[†‡] **Shih-Hung Liu**[*‡] **Cheng-Lung Sung**[*‡] **Wen-Lian Hsu**[†‡]

[†]Department of Computer Science, National Tsing-Hua University

[*]Department of Electrical Engineering, National Taiwan University

[‡]Institute of Information Science, Academia Sinica

{tmjiang,journey,clsung,hsu}@iis.sinica.edu.tw

Abstract

This paper proposes a novel feature for conditional random field (CRF) model in Chinese word segmentation system. The system uses a conditional random field as machine learning model with one simple feature called *term contributed boundaries* (TCB) in addition to the "BIEO" character-based label scheme. TCB can be extracted from unlabeled corpora automatically, and segmentation variations of different domains are expected to be reflected implicitly. The dataset used in this paper is the closed training task in CIPS-SIGHAN-2010 bakeoff, including simplified and traditional Chinese texts. The experiment result shows that TCB does improve "BIEO" tagging domain-independently about 1% of the F1 measure score.

Keywords: Term contributed boundary, Conditional Random fields, Chinese word segmentation.

1. Introduction

Word segmentation is a trivial problem for most Western language, since there are clear delimiters (e.g. spaces) for individual words. However, for some Asia languages such as Chinese, Japanese and other language do not have word delimiters, word segmentation problem will be encountered if we want to do some further language processing, e.g. information retrieval, summarization and so on. Thus, the Chinese word segmentation could be viewed as a fundamental problem for natural language processing.

Chinese word segmentation is still a challenging issue, and there is contest held in SIGHAN community [1]. The CIPS-SIGHAN-2010 bakeoff task of Chinese word segmentation is focused on cross-domain texts [2]. The design of data set is challenging particularly. The domain-specific training corpora remain unlabeled, and two of the test corpora keep domains unknown before releasing, therefore it is not easy to apply ordinary machine learning approaches, especially for the closed training evaluations.

Proceedings of the 22nd Conference on Computational Linguistics and Speech Processing (ROCLING 2010), Pages 143-156,
Puli, Nantou,Taiwan, September 2010.

Traditional approach for Chinese word segmentation problem is adopted dictionary along with lots of rules to segment the unlabelled texts [3]. Recent years, the statistical machine learning models, such as Hidden Markov Model (HMM) [4], Maximum Entropy Markov Model (MEMM) [5] and Conditional Random Field (CRF) [6], show the moderate performance for sequential labeling problem, especially CRF achieves better outcome.

In this paper we propose a novel feature named term contributed boundary (TCB) for CRF model training. Since term contributed boundary extraction [10] is unsupervised, it is suitable for closed training task that any external resource or extra knowledge is not allowed. Without proper knowledge, the closed task of word segmentation can be hard when out-of-vocabulary (OOV) sequences occurred, where TCB extracted from test data directly may help.

We also compare different character based label scheme “BI”, “BIO” and “BIEO” for model training. “B,” “I,” “E” and “O” mean the beginning of word, the internal of word, the end of word and the single character word, respectively. The character-based “BIO” tagging of Conditional Random Field has been widely used in Chinese word segmentation recently [11, 12, 13]. From the experiments, “BIEO” labeling shows the better performance than “BI” and “BIO”.

The layout of this paper is as follows. We briefly introduce of CRF in Section 2. The novel feature term contributed boundary will be given in Section 3. Section 4 describes the data set and experimental results with error analysis. The conclusion is in Section 5.

2. Conditional Random Fields

Conditional random fields (CRF) are undirected graphical models trained to maximize a conditional probability of random variables X and Y, and the concept is well established for sequential labeling problem [6]. Given an input sequence (or observation sequence) $\mathrm{X} = x_1 \dots x_T$ and label sequence $\mathrm{Y} = y_1 \dots y_T$, a conditional probability of linear-chain CRF with parameters $\Lambda = \{\lambda_1, ..., \lambda_n\}$ can be defined as:

$$\mathrm{P}_\lambda(\mathrm{Y} \mid \mathrm{X}) = \frac{1}{\mathrm{Z_X}} \exp\left(\sum_{t=1}^{T} \sum_{k} \lambda_k f_k(y_{t-1}, y_t, X, t) \right) \tag{1}$$

where $\mathrm{Z_X}$ is the normalization constant that makes probability of all label sequences sum to one, $f_k(y_{t-1}, y_t, X, t)$ is a feature function which is often binary valued, but can be real valued, and λ_k is a learned weight associated with feature f_k.

The feature functions can measure any aspect of state transition $y_{t-1} \to y_t$, and the entire observation sequence X, centered at the current position t. For instance, one feature function might be value one when y_{t-1} is the state B, y_t is the state I, and x_t is the character “全”.

Given such a model as defined in Eq. 1, the most probable labeling sequence for an input sequence X is as follow.

$$y^* = \arg\max_{Y} P_\Lambda(Y \mid X) \tag{2}$$

Eq. 2 can be efficiently calculated by dynamic programming using Viterbi algorithm. The more details about concepts of CRF and learning parameters could be refer to [7]. Figure 1 shows the CRF tagging, which is based on BIEO label training, in test phase when given the un-segmented input.

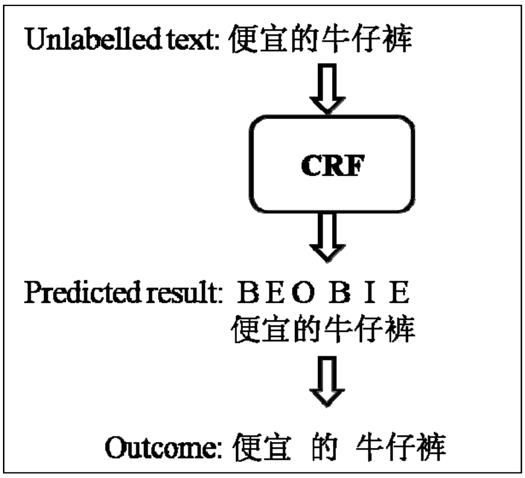


Figure 1. Illustration of CRF prediction

3. Term Contributed Boundary

The word boundary and the word frequency are the standard notions of frequency in corpus-based natural language processing, but they lack the correct information about the actual boundary and frequency of a phrase’s occurrence. The distortion of phrase boundaries and frequencies was first observed in the Vodis Corpus when the bigram “RAIL ENQUIRIES” and trigram “BRITISH RAIL ENQUIRIES” were examined and reported by O'Boyle [8]. Both of them occur 73 times, which is a large number for such a small corpus. “ENQUIRIES” follows “RAIL” with a very high probability when it is preceded by “BRITISH.” However, when “RAIL” is preceded by words other than “BRITISH,” “ENQUIRIES” does not occur, but words like “TICKET” or “JOURNEY” may. Thus, the bigram “RAIL ENQUIRIES” gives a misleading probability that “RAIL” is followed by “ENQUIRIES” irrespective of what precedes it. This problem happens not only with word-token corpora but also with corpora in which all the compounds are tagged as units since overlapping N-grams still appear, therefore corresponding solutions such as Zhang *et al.* were proposed [9].

We uses suffix array algorithm to calculate exact boundaries of phrase and their frequencies [10], called *term contributed boundaries* (TCB) and *term contributed frequencies* (TCF), respectively, to analogize similarities and differences with the *term frequencies* (TF). For example, in Vodis Corpus, the original TF of the term “RAIL ENQUIRIES” is 73. However, the actual TCF of “RAIL ENQUIRIES” is 0, since all of the frequency values are contributed by the term “BRITISH RAIL ENQUIRIES”. In this case, we can see that ‘BRITISH RAIL ENQUIRIES’ is really a more frequent term in the corpus, where “RAIL ENQUIRIES” is not. Hence the TCB of “BRITISH RAIL ENQUIRIES” is ready for CRF tagging as “BRITISH/TB RAIL/TI ENQUIRIES/TI,” for example, “TB” means beginning of the term contributed boundary and “TI” is the other place of term contributed boundary.

In Chinese, similar problems occurred as Lin and Yu reported [14, 15], consider the following Chinese text:

“自然科學的重要” (the importance of natural science) and

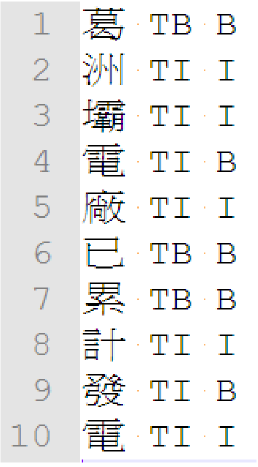
“自然科學的研究是唯一的途徑” (the research on natural science is the only way).

In the above text, there are many string patterns that appear more than once. Some patterns are listed as follows:

“自然科學” (natural science) and

“自然科學的” (of natural science).

They suggested that it is very unlikely that a random meaningless string will appear more than once in a corpus. The main idea behind our proposed method is that if a Chinese string pattern appears two or more times in the text, then it may be useful. However, not all the patterns which appear two or more times are useful. In the above text, the pattern “然科” has no meaning. Therefore they proposed a method that is divided into two steps. The first step is to search through all the characters in the corpus to find patterns that appear more than once. Such patterns are gathered into a database called *MayBe*, which means these patterns may be “Chinese Frequent Strings” as they defined. The entries in *MayBe* consist of strings and their numbers of occurrences. The second step is to find the net frequency of occurrence for each entry in the above database. The net frequency of occurrence of an entry is the number of appearances which do not depend on other super-strings. For example, if the content of a text is “自然科學，自然科學” (natural science, natural science), then the net frequency of occurrence of “自然科學” is 2, and the net frequency of occurrence of “自然科” is zero since the string “自然科” is brought about by the string “自然科學.” They exclude the appearances of patterns which are brought about by others; hence their method is actually equivalent to the suffix array algorithm we apply for calculating TCB and TCF, and the annotated input string for CRF will be “自/TB 然/TI 科/TI 學/TI”. The Figure 2 demonstrates one labeled phrase from the training data.



```
葛 TB B
洲 TI I
壩 TI I
電 TI B
廠 TI I
已 TB B
累 TB B
計 TI I
發 TI B
電 TI I
```

Figure 2. Example of training data for CRF with “BI” and TCB

4. Experiments

4.1 Dataset

The corpora used in this paper are CIPS-SIGHAN-2010 bakeoff dataset which contain simplified Chinese (SC) and traditional Chinese (TC) texts. There are two types of training corpus for each language, including labeled training data (Chinese text which has been segmented into words) and unlabelled training data. The unlabeled corpus used in this bake-off task covers two domains: literature and computer science. The corpus for each domain is a pure text file which contains about 100,000 Chinese characters. The test corpus covers four domains, two of which are literature (denoted as Test-A) and computer science (denoted as Test-B), and the other two domains are medicine (denoted as Test-C) and finance (denoted as Test-D).

4.2 Experimental Results

The evaluation metric of word segmentation task is precision (P), recall (R), F1 measure (F) and OOV recall (OOV) which are list as follow.

- Precision = $\dfrac{\text{the number of words that are correctly segmented}}{\text{the number of words that are segmented}} \times 100\%$
- Recall = $\dfrac{\text{the number of words that are correctly segmented}}{\text{the number of words in the reference}} \times 100\%$
- F1 measure = $2 \times P \times R / (P + R)$
- OOV Recall = $\dfrac{\text{the number of OOV words that are correctly segmented}}{\text{the number of OOV words in the reference}} \times 100\%$

4.2.1 Experiments of Comparison with BI, BIO and BIEO

Experiments here evaluate the performance between three different label schemes "BI", "BIO" and "BIEO" for two types (SC and TC) in four domains (Test-A, Test-B, Test-C and Test-D). The result shows in Table 1. The scheme "BIEO" outperforms "BI" and "BIEO" on F1 measure, except at SC-Test-B. The domain B is computer science and its test data mingles many English words. In the end of section 4.2.2, we will deal with this problem using post-processing for English words.

		R	P	F	OOV
SC-Test-A	BIEO	**0.913**	**0.921**	**0.917**	**0.554**
	BIO	0.906	0.916	0.911	0.539
	BI	0.896	0.907	0.901	0.508
SC-Test-B	BIEO	**0.868**	0.785	0.824	0.379
	BIO	**0.868**	**0.797**	**0.831**	**0.410**
	BI	0.850	0.763	0.805	0.327
SC-Test-C	BIEO	**0.903**	**0.899**	**0.901**	**0.602**
	BIO	0.897	0.897	0.897	0.590
	BI	0.888	0.886	0.887	0.551
SC-Test-D	BIEO	**0.914**	**0.908**	**0.911**	**0.516**
	BIO	0.900	0.903	0.901	0.472
	BI	0.888	0.891	0.890	0.419
TC-Test-A	BIEO	**0.879**	**0.905**	**0.892**	0.725
	BIO	0.873	0.898	0.886	**0.727**
	BI	0.856	0.884	0.870	0.674
TC-Test-B	BIEO	**0.909**	**0.937**	**0.923**	0.568
	BIO	0.906	0.932	0.919	**0.578**
	BI	0.894	0.920	0.907	0.551
TC-Test-C	BIEO	**0.910**	**0.929**	**0.920**	0.716
	BIO	0.902	0.923	0.913	**0.722**
	BI	0.891	0.914	0.902	0.674
TC-Test-D	BIEO	**0.928**	**0.939**	**0.933**	0.761
	BIO	0.924	0.934	0.929	**0.765**
	BI	0.908	0.922	0.915	0.722

Table 1. Comparison of BI, BIO and BIEO

4.2.2 Term Contributed Boundary Experiments with BI as Baseline

In this section, we evaluate the performance of term contributed boundary as a feature in CRF model training. The label scheme “BI” of ground truth has been treated as baseline for comparison with TCB features, which label scheme is also “BI”. There are several different experiments that we have done which are showed in Table 2 and Table 3a and Table 3b. The configuration is about the trade-off between data sparseness and domain fitness. For the sake of OOV issue, TCBs from all the training and test corpora are included in the configuration of results. For potentially better consistency to different types of text, TCBs from the training corpora and/or test corpora are grouped by corresponding domains of test corpora. Table 2, Table 3a and Table 3b provide the details, where the baseline is the character-based “BI” tagging, and others are “BI” with additional different TCB configurations: TCB$_{\text{all}}$ stands for the TCB extracted from all training data and all test data; TCB$_{\text{a}}$, TCB$_{\text{b}}$, TCB$_{\text{ta}}$, TCB$_{\text{tb}}$, TCB$_{\text{tc}}$, TCB$_{\text{td}}$ represents TCB extracted from the training corpus A, B, and the test corpus A, B, C, D, respectively.

		R	P	F	OOV
SC-Test-A	BI	0.896	0.907	0.901	0.508
	TCB$_{\text{all}}$	**0.917**	**0.921**	**0.919**	**0.699**
SC-Test-B	BI	0.850	0.763	0.805	0.327
	TCB$_{\text{all}}$	**0.876**	**0.799**	**0.836**	**0.456**
SC-Test-C	BI	0.888	0.886	0.887	0.551
	TCB$_{\text{all}}$	**0.900**	**0.896**	**0.898**	**0.699**
SC-Test-D	BI	0.888	0.891	0.890	0.419
	TCB$_{\text{all}}$	**0.910**	**0.906**	**0.908**	**0.562**
TC-Test-A	BI	0.856	0.884	0.870	**0.674**
	TCB$_{\text{all}}$	**0.871**	**0.891**	**0.881**	0.670
TC-Test-B	BI	0.894	0.920	0.907	0.551
	TCB$_{\text{all}}$	**0.913**	**0.917**	**0.915**	**0.663**
TC-Test-C	BI	0.891	0.914	0.902	**0.674**
	TCB$_{\text{all}}$	**0.900**	**0.915**	**0.908**	0.668
TC-Test-D	BI	0.908	0.922	0.915	0.722
	TCB$_{\text{all}}$	**0.929**	0.922	**0.925**	**0.732**

Table 2. Baseline vs. Term Contributed Boundary Results

		F	OOV
SC-Test-A	TCB$_{ta}$	0.918	0.690
	TCB$_{a}$	0.917	0.679
	TCB$_{ta}$ + TCB$_{a}$	0.917	0.690
	TCB$_{all}$	**0.919**	**0.699**
SC-Test-B	TCB$_{tb}$	0.832	**0.465**
	TCB$_{b}$	0.828	0.453
	TCB$_{tb}$ + TCB$_{b}$	0.830	0.459
	TCB$_{all}$	**0.836**	0.456
SC-Test-C	TCB$_{tc}$	0.897	0.618
	TCB$_{all}$	**0.898**	**0.699**
SC-Test-D	TCB$_{td}$	0.905	0.557
	TCB$_{all}$	**0.910**	**0.562**

Table 3a. Simplified Chinese Domain-specific TCB vs. TCB$_{all}$

		F	OOV
TC-Test-A	TCB$_{ta}$	**0.889**	0.706
	TCB$_{a}$	0.888	0.690
	TCB$_{ta}$ + TCB$_{a}$	**0.889**	**0.710**
	TCB$_{all}$	0.881	0.670
TC-Test-B	TCB$_{tb}$	0.911	0.636
	TCB$_{b}$	**0.921**	**0.696**
	TCB$_{tb}$ + TCB$_{b}$	0.912	0.641
	TCB$_{all}$	0.915	0.663
TC-Test-C	TCB$_{tc}$	**0.918**	**0.705**
	TCB$_{all}$	0.908	0.668
TC-Test-D	TCB$_{td}$	**0.927**	0.717
	TCB$_{all}$	0.925	**0.732**

Table 3b. Traditional Chinese Domain-specific TCB vs. TCB$_{all}$

Table 2 indicates that F1 measure scores can be improved by TCB about 1%, domain-independently. Table 3a and Table 3b give a hint of the major contribution of performance has been done by TCB extracted from each test corpus.

In order to deal with English words, we apply post-processing to the segmented data. It simply recovers alphanumeric sequences according to their original segments in the training data. Table 4 shows the experiment result after post-processing. The performance has been improved, especially on the domain B of computer science, since its data consists of a lot of technical terms in English.

		F1 measure score	
		Before	After
SC-A	BIO	0.911	0.918
	BI	0.901	0.908
	TCB_{ta}	0.918	0.920
	$\text{TCB}_{ta\,+}\,\text{TCB}_{a}$	0.917	0.920
	TCB_{all}	0.919	**0.921**
SC-B	BIO	0.831	**0.920**
	BI	0.805	0.910
	TCB_{tb}	0.832	0.917
	$\text{TCB}_{tb\,+}\,\text{TCB}_{b}$	0.830	0.916
	TCB_{all}	0.836	0.916
SC-C	BIO	0.897	**0.904**
	BI	0.887	0.896
	TCB_{tc}	0.897	0.901
	TCB_{all}	0.898	0.902
SC-D	BIO	0.901	**0.919**
	BI	0.890	0.908
	TCB_{td}	0.905	0.915
	TCB_{all}	0.908	0.918

Table 4. F1 scores before and after the English problem fixed

4.2.3 Term Contributed Boundary Experiments with BIO and BIEO

In this section we combine the TCB feature with “BIEO” to compare with “BIO”. Table 5a and Table 5b show the experimental results. We find that our TCB feature is robustness, which would not affected by different label scheme. This meets our conjecture, and the experiments are fit our expectations.

		F	OOV
SC-Test-A	BIEO, TCB_a	**0.931**	0.720
	BIEO, TCB_{all}	**0.931**	**0.723**
	BIO, TCB_a	0.926	0.717
	BIO, TCB_{all}	0.925	0.711
SC-Test-B	BIEO, TCB_b	0.840	0.473
	BIEO, TCB_{all}	0.833	0.451
	BIO, TCB_b	**0.849**	**0.512**
	BIO, TCB_{all}	0.840	0.472
SC-Test-C	BIEO, TCB_c	0.911	**0.651**
	BIEO, TCB_{all}	**0.912**	0.636
	BIO, TCB_c	0.907	0.646
	BIO, TCB_{all}	0.907	0.628
SC-Test-D	BIEO, TCB_d	**0.923**	**0.631**
	BIEO, TCB_{all}	**0.923**	0.613
	BIO, TCB_d	0.916	0.605
	BIO, TCB_{all}	0.918	0.597

Table 5a. Simplified Chinese Domain-specific TCB vs. TCB_{all} with BIO and BIEO

		F	OOV
TC-Test-A	BIEO, TCB_a	**0.909**	**0.747**
	BIEO, TCB_{all}	0.908	0.743
	BIO, TCB_a	0.904	0.744
	BIO, TCB_{all}	0.906	0.744
TC-Test-B	BIEO, TCB_b	0.943	0.771
	BIEO, TCB_{all}	0.940	0.754
	BIO, TCB_b	**0.945**	**0.804**
	BIO, TCB_{all}	**0.945**	**0.804**
TC-Test-C	BIEO, TCB_c	**0.931**	0.737
	BIEO, TCB_{all}	0.930	0.730
	BIO, TCB_c	0.928	0.743
	BIO, TCB_{all}	0.929	**0.745**
TC-Test-D	BIEO, TCB_d	0.942	0.768
	BIEO, TCB_{all}	0.943	0.771
	BIO, TCB_d	**0.944**	**0.778**
	BIO, TCB_{all}	0.943	0.777

Table 5b. Traditional Chinese Domain-specific TCB vs. TCB_{all} with BIO and BIEO

For the sake of consistency, we do extra experiments using label schemes either “BIEO” or “BIO” to label TCB features, and denote them as TE-TCB and TO-TCB. In these schemes, “TB,” “TI,” “TE” and “TO” are tags for the head of TCB, the middle of TCB, the tail of TCB, and the single character word of TCB, respectively. TE-TCB uses all tags but TO-TCB excludes the tag “TE.” Table 6a and Table 6b show the comparisons between the original TCB, TO-TCB and TE-TCB. The result suggests that TO-TCB and TE-TCB may not have stable and significant improvements to the original TCB scheme that consists of only “TB” and “TI.” We suspect that it is because single character words of TCB and the tail character of TCB sometimes conflict with the word boundaries of gold standard, after all the concept of TCB is from suffix pattern, not from linguistic design.

		F	OOV
SC-Test-A	BIEO, TCB$_a$	0.931	0.720
	BIEO, TO- TCB$_a$	0.929	0.719
	BIEO, TE-TCB$_a$	**0.932**	0.719
	BIEO, TCB$_{all}$	0.931	**0.723**
	BIEO, TO-TCB$_{all}$	0.929	0.719
	BIEO, TE-TCB$_{all}$	0.931	0.719
SC-Test-B	BIEO, TCB$_b$	0.840	0.473
	BIEO, TO-TCB$_b$	**0.841**	0.473
	BIEO, TE-TCB$_b$	0.838	**0.475**
	BIEO, TCB$_{all}$	0.833	0.451
	BIEO, TO-TCB$_{all}$	0.835	0.443
	BIEO, TE-TCB$_{all}$	0.838	0.455
SC-Test-C	BIEO, TCB$_c$	0.911	0.651
	BIEO, TO-TCB$_c$	0.910	0.655
	BIEO, TE-TCB$_c$	**0.913**	**0.665**
	BIEO, TCB$_{all}$	0.912	0.636
	BIEO, TO-TCB$_{all}$	0.906	0.599
	BIEO, TE-TCB$_{all}$	0.909	0.625
SC-Test-D	BIEO, TCB$_d$	0.923	0.631
	BIEO, TO-TCB$_d$	0.916	0.605
	BIEO, TE-TCB$_d$	**0.925**	**0.643**
	BIEO, TCB$_{all}$	0.923	0.613
	BIEO, TO-TCB$_{all}$	0.921	0.592
	BIEO, TE-TCB$_{all}$	0.923	0.612

Table 6a. Comparisons between TCB, TO-TCB and TE-TCB for Simplified Chinese test set

		F	OOV
TC-Test-A	BIEO, TCB_a	**0.909**	**0.747**
	BIEO, TO-TCB_a	0.905	0.732
	BIEO, TE-TCB_a	0.907	0.733
	BIEO, TCB_{all}	0.908	0.743
	BIEO, TO-TCB_{all}	0.905	0.733
	BIEO, TE-TCB_{all}	0.906	0.731
TC-Test-B	BIEO, TCB_b	**0.943**	**0.771**
	BIEO, TO-TCB_b	0.935	0.734
	BIEO, TE-TCB_b	0.941	0.759
	BIEO, TCB_{all}	0.940	0.754
	BIEO, TO-TCB_{all}	0.935	0.732
	BIEO, TE-TCB_{all}	0.940	0.754
TC-Test-C	BIEO, TCB_c	0.931	**0.737**
	BIEO, TO-TCB_c	0.930	0.722
	BIEO, TE-TCB_c	0.931	0.731
	BIEO, TCB_{all}	0.930	0.730
	BIEO, TO-TCB_{all}	0.927	0.713
	BIEO, TE-TCB_{all}	**0.932**	0.730
TC-Test-D	BIEO, TCB_d	0.942	0.768
	BIEO, TO-TCB_d	0.939	0.758
	BIEO, TE-TCB_d	**0.944**	0.769
	BIEO, TCB_{all}	0.943	0.771
	BIEO, TO-TCB_{all}	0.939	0.759
	BIEO, TE-TCB_{all}	**0.944**	**0.779**

Table 6b. Comparisons between TCB and TX-TCB for Traditional Chinese test set

4.3 Error Analysis

The most significant type of error in our results is unintentionally segmented English words. Rather than developing another set of tag for English alphabets, we applies post-processing to fix this problem under the restriction of closed training by using only alphanumeric character information. Table 4 compares F1 measure score of the Simplified Chinese experiment results before and after the post-processing.

The major difference between gold standards of the Simplified Chinese corpora and the Traditional Chinese corpora is about non-Chinese characters. All of the alphanumeric and the punctuation sequences are separated from Chinese sequences in the Simplified Chinese corpora, but can be part of the Chinese word segments in the Traditional Chinese corpora.

For example, a phrase “服用/simvastatin/（/statins 類/的/一/種/）,” where ‘/’ represents the word boundary, from the domain C of the test data cannot be either recognized by “BIEO” and/or TCB tagging approaches, or post-processed. This is the reason why Table 4 does not come along with Traditional Chinese experiment results.

Some errors are due to inconsistencies in the gold standard of non-Chinese character, For example, in the Traditional Chinese corpora, some percentage digits are separated from their percentage signs, meanwhile those percentage signs are connected to parentheses right next to them.

5. Conclusions

This paper introduces a simple CRF feature called term contributed boundaries (TCB) for Chinese word segmentation. The experiment result shows that it can improve the basic “BIEO” tagging scheme about 1% of the F1 measure score, domain-independently.

Further tagging scheme for non-Chinese characters are desired for recognizing some sophisticated gold standard of Chinese word segmentation that concatenates alphanumeric characters to Chinese characters.

Acknowledgement

The CRF model used in this paper is developed based on CRF++ version 0.49, http://crfpp.sourceforge.net/

The TCBs and TCFs are calculated by YASA (Yet Another Suffix Array) version 0.2.3, http://yasa.newzilla.org/

6. References

[1] SIGHAN, http://sighan.cs.uchicago.edu/

[2] CIPS-SIGHAN-2010, http://www.cipsc.org.cn/clp2010/

[3] Ma, Wei-Yun and Keh-Jiann Chen, “Introduction to CKIP Chinese Word Segmentation System for the First International Chinese Word Segmentation Bakeoff,” in *Proceedings of ACL, Second SIGHAN Workshop on Chinese Language Processing*, pp. 168–171, 2003.

[4] Lawrence R. Rabiner, “A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition,” in *Proceedings of the IEEE*, Vol. **77,** No. 2, pp. 257–286, 1989.

[5] McCallum, A., Freitag, D. & Pereira, F., “Maximum Entropy Markov Models for

Information Extraction and Segmentation," in *Proceedings of ICML*, 2000.

[6] John Lafferty, Andrew McCallum, and Fernando Pereira, "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data," in *Proceedings of ICML*, pp. 591–598, 2001.

[7] Hanna M. Wallach, "Conditional Random Fields: An Introduction," *Technical Report MS-CIS-04-21*, Department of Computer and Information Science, University of Pennsylvania, 2004.

[8] Peter O'Boyle, *A Study of an N-Gram Language Model for Speech Recognition*, PhD thesis, Queen's University Belfast, 1993

[9] Ruiqiang Zhang, Genichiro Kikui, and Eiichiro Sumita, "Subword-Based Tagging by Conditional Random Fields for Chinese Word Segmentation," in *Proceedings of the Human Language Technology Conference of the NAACL*, pp. 193–196, New York, USA, 2006.

[10] Cheng-Lung Sung, Hsu-Chun Yen, and Wen-Lian Hsu, "Compute the Term Contributed Frequency," in *Proceedings of the 2008 Eighth International Conference on Intelligent Systems Design and Applications*, pp. 325–328, Washington, D.C., USA, 2008.

[11] Fuchun Peng and Andrew McCallum, "Chinese segmentation and new word detection using conditional random fields," in *Proceedings of Coling-2004*, pp. 562–568, Geneva, Switzerland, 2004

[12] Huihsin Tseng, Pichuan Chang, Galen Andrew, Daniel Jurafsky, and Christopher Manning, "A Conditional Random Field Word Segmenter for SIGHAN Bakeoff 2005," in *Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing*, Jeju, Korea, 2005.

[13] Nianwen Xue and Libin Shen, "Chinese Word Segmentation as LMR Tagging," in *Proceedings of the Second SIGHAN Workshop on Chinese Language Processing*, 2003.

[14] Yih-Jeng Lin and Ming-Shing Yu, "Extracting Chinese Frequent Strings without a Dictionary from a Chinese Corpus and its Applications," *Journal of Information Science and Engineering*, Vol. 17, pp. 805–824, 2001.

[15] Yih-Jeng Lin and Ming-Shing Yu, "The Properties and Further Applications of Chinese Frequent Strings," *Computational Linguistics and Chinese Language Processing*, Vol. 9, No. 1, pp. 113–128, February 2004.