Term Contributed Boundary Tagging by Conditional Random Fields for SIGHAN 2010 Chinese Word Segmentation Bakeoff

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 presents a Chinese word segmentation system submitted to the closed training evaluations of CIPS-SIGHAN-2010 bakeoff. The system uses a conditional random field model with one simple feature called *term contributed boundaries* (TCB) in addition to the "BI" character-based tagging approach. TCB can be extracted from unlabeled corpora automatically, and segmentation variations of different domains are expected to be reflected implicitly. The experiment result shows that TCB does improve "BI" tagging domain-independently about 1% of the F1 measure score.

1 Introduction

The CIPS-SIGHAN-2010 bakeoff task of Chinese word segmentation is focused on cross-domain texts. 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.

2 Methodology

2.1 The "BI" Character-Based Tagging of Conditional Random Field as Baseline

The character-based "OBI" tagging of Conditional Random Field (Lafferty et al., 2001) has been widely used in Chinese word segmentation recently (Xue and Shen, 2003; Peng and McCallum, 2004; Tseng et al., 2005). Under the scheme, each character of a word is labeled as 'B' if it is the first character of a multiple-character word, or 'I' otherwise. If the character is a single-character word itself, "O" will be its label. As Table 1 shows, the lost of performance is about 1% by replacing "O" with "B" for character-based CRF tagging on the dataset of CIPS-SIGHAN-2010 bakeoff task of Chinese word segmentation, thus we choose "BI" as our baseline for simplicity, with this 1% lost bearing in mind. In tables of this paper, SC stands for Simplified Chinese and TC represents for Traditional Chinese. Test corpora of SC and TC are divided into four domains, where suffix A, B, C and D attached, for texts of literature, computer, medicine and finance, respectively.

		R	P	F	OOV
SC-A	OBI	0.906	0.916	0.911	0.539
	BI	0.896	0.907	0.901	0.508
SC-B	OBI	0.868	0.797	0.831	0.410
	BI	0.850	0.763	0.805	0.327
SC-C	OBI	0.897	0.897	0.897	0.590
	BI	0.888	0.886	0.887	0.551
SC-D	OBI	0.900	0.903	0.901	0.472
	BI	0.888	0.891	0.890	0.419
TC-A	OBI	0.873	0.898	0.886	0.727
	BI	0.856	0.884	0.870	0.674
TC-B	OBI	0.906	0.932	0.919	0.578
	BI	0.894	0.920	0.907	0.551
TC-C	OBI	0.902	0.923	0.913	0.722
	BI	0.891	0.914	0.902	0.674
TC-D	OBI	0.924	0.934	0.929	0.765
	BI	0.908	0.922	0.915	0.722

Table 1. OBI vs. BI; where the lost of F > 1%, such as SC-B, is caused by incorrect English segments that will be discussed in the section 4.

2.2 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 (1993). 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 those of Zhang et al. (2006) were proposed.

We uses suffix array algorithm to calculate exact boundaries of phrase and their frequencies (Sung et al., 2008), 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 EN QUIRIES". 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/TB ENQUIRIES/TI," for example.

3 Experiments

Besides submitted results, there are several different experiments that we have done. 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 submitted 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 and Table 3 provide the details, where the baseline is the character-based "BI" tagging, and others are "BI" with additional different TCB configurations: TCB_{all} stands for the submitted results; TCB_a, TCB_b, TCB_{ta}, TCB_{tb}, TCB_{tc}, TCB_{td} represents TCB extracted from the training corpus A, B, and the test corpus A, B, C, D, respectively. Table 2 indicates that F1 measure scores can be improved by TCB about 1%, domain-independently. Table 3 gives a hint of the major contribution of performance is from TCB of each test corpus.

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

Table 2. Baseline vs. Submitted Results

		F	OOV
SC-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-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-C	TCB_{tc}	0.897	0.618
	TCB_{all}	0.898	0.699
SC-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-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-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-C	TCB_{tc}	0.918	0.705
	TCB_{all}	0.908	0.668
TC-D	TCB_{td}	0.927	0.717
	TCB_{all}	0.925	0.732

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

4 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.

		F1 measure score	
		before	after
SC-A	OBI	0.911	0.918
	BI	0.901	0.908
	TCB_{ta}	0.918	0.920
	$TCB_{ta} + TCB_{a}$	0.917	0.920
	TCB_{all}	0.919	0.921
SC-B	OBI	0.831	0.920
	BI	0.805	0.910
	TCB_{tb}	0.832	0.917
	$TCB_{tb} + TCB_{b}$	0.830	0.916
	TCB_{all}	0.836	0.916
SC-C	OBI	0.897	0.904
	BI	0.887	0.896
	TCB_{tc}	0.897	0.901
	TCB_{all}	0.898	0.902
SC-D	OBI	0.901	0.919
	BI	0.890	0.908
	TCB_{td}	0.905	0.915
	TCB_{all}	0.908	0.918

Table 4. F1 measure scores before and after English Problem Fixed

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 / ）" ('/' represents the word boundary) from the domain C of the test data cannot be either recognized by "BI" 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 Conclusion

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 "BI" 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++, http://crfpp.sourceforge.net/

Term Contributed Boundaries used in this paper are extracted by YASA, http://yasa.newzilla.org/

References

John Lafferty, Andrew McCallum, and Fernando Pereira. 2001. Conditional random fields: probabilistic models for segmenting and labeling sequence data. In *Proceedings of International Conference of Machine Learning*, 591–598.

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

Fuchun Peng and Andrew McCallum. 2004. Chinese segmentation and new word detection using conditional random fields. In *Proceedings of International Conference of Computational linguistics,* 562–568, Geneva, Switzerland.

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

Huihsin Tseng, Pichuan Chang, Galen Andrew, Daniel Jurafsky, and Christopher Manning. 2005. A conditional random field word segmenter for Sighan bakeoff 2005. In *Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing*, Jeju, Korea.

Nianwen Xue and Libin Shen. 2003. Chinese word-segmentation as LMR tagging. In *Proceedings of the Second SIGHAN Workshop on Chinese Language Processing*.

Ruiqiang Zhang, Genichiro Kikui, and Eiichiro Sumita. 2006. Subword-based tagging by conditional random fields for Chinese word segmentation. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, 193-196, New York, USA.