A Feature-Rich CRF Segmenter for Chinese Micro-blog

Yabin Leng(✉), Weiwei Liu, Sheng Wang, and Xiaojie Wang

School of Computer Science,
Beijing University of Posts and Telecommunications,
China, 100876
{lengyabin, liuww_victor, bupt10211677, xjwang}@bupt.edu.cn

Abstract. This paper describes our system for Chinese word segmentation of micro-blog text, one of the NLPCC-ICCPOL 2016 Shared Tasks [1]. The CRF (Conditional Random Field) model is employed to model word segmentation as a sequence labeling problem, 7 sets of features are selected to train the CRF model. The system achieves f_b 0.798144 on closed track, 0.81968 on semi-open track, and 0.82217 on open track with weighted measures [2].

Keywords: Chinese word segmentation on Micro-blog · Sequence labeling · CRF

1 Introduction

Chinese word segmentation is the fundamental task of Chinese natural language processing [3]. Lots of models have been proposed for the task, such as ME (Maximum Entropy) [4], CRF (Conditional Random Field) [5] to deep learning such as LSTM (Long Short-Term Memory) [6]. The performance of traditional texts segmentation has been improved significantly. However, the results of previous methods on micro-blog is not as good as those in traditional texts. One of the main reasons is that micro-blog shows a very different wording style with traditional texts such as newspapers and radio reports. Sentences in micro-blog contain many new words, for instance, "木有", "棒棒哒". Not only the number of new words increases rapidly, but also they are constructed in very different way from those new words in traditional texts. It brings a big challenge for Chinese word segmentation.

In this paper, we take full advantages of both unsupervised features and supervised features to discover new words from unlabeled dataset. We explore some features especially for micro-blog, such as reduplication feature. The new words recognition is significantly improved with those features.

The remainders of this paper are organized as follows. In Section 2, we introduce the model and features. In Section 3, we show our experimental results and analysis on the evaluation data. Finally, we conclude the paper and discuss future work in Section 4.

2 Our Method

We introduce the method in this section. There are mainly two parts including the model and the features.

2.1 Model

Chinese word segmentations is often modeled as sequence labeling on Chinese characters. Considering the great success that CRF has achieved in sequence labeling and its capability of making use of different features, we use CRF as basic model in our method.

A linear-chain CRF with parameters $\Lambda = \{\lambda_1, ...\}$ defines a conditional probability for a label sequence y = $y_1...y_n$ given an input sequence x = $x_1...x_n$ to be

$$p_\Lambda(y|x) = \frac{exp(\sum_{i=1}^{n}\sum_k \lambda_k f_k(y_{i-1}, y_i, x, i))}{Z(x)} \quad (1)$$

where Z(x) is the normalization factor, $f_k(y_{i-1}, y_i, x, i)$ is a feature function which is often binary-valued, but can be real-valued, λ_k is a learned weight associated with feature f_k.

A 6-tag set including B, B_2, B_3, M, E and S is used in this paper to label the position of characters in words. 6-tag set has been shown better performance than other tag sets in Zhao [7], Table 1 shows how to label the characters in words with different lengths.

Table 1: Tagging for a word

Length of word	Tags for a word
1	S
2	BE
3	BB_2E
4	BB_2B_3E
5	BB_2B_3ME
>5	BB_2B_3M...E

2.2 Features

Feature selection has a great influence on the performance of CRF model. Follows are some features used in our CRF model. Feature templates we designed for all features below (except AV) are shown in Table 2.

Basic Features There are three types of basic features, including Character Feature, Character Type Feature, and Reduplication Feature.

Character Feature (CF): 6 n-gram character feature are used, the templates are listed in Table 2.

Table 2: The templates of CF

Feature	Instruction
C_{-1}	the previous character
C_0	the current character
C_1	the next character
$C_{-1}C_0$	the previous character and the current character
C_0C_1	the current character and the next character
$C_{-1}C_1$	the previous character and the next character

Character Type Feature (CTF): The characters are divided into 7 categories, including numbers (0-9), unit of measurement, punctuations, English characters (a-zA-Z), Chinese numbers (零 (zeor) - 十 (ten)), Chinese characters and other characters.

Reduplication Feature (RF): Reduplication feature is special for micro-blog. There are many words with reduplication forms like "AAB", "ABAB", "AABB" or "AAA" in micro-blog. For example, "棒棒哒", "就是就是", "开开心心". There are also some long words consisting of punctuation marks or Chinese modal particle such as "哈" and "嘻". In order to better deal with these kinds of words, we propose reduplication feature shown in Table 3.

Table 3: The description of RF

type	Instruction
3	shape like "AAA"
2	shape like "ABA"
1	shape like "ABB"
0	others

Conditional Entropy Feature (CEF) Conditional entropy feature is a kind of unsupervised feature. Gao's [9] experiments proved that conditional entropy feature improves the performance of word segmentation model. Given a character C, the forward and backward conditional entropies are calculated by formula (2) and (3) respectively. And continuous conditional entropies are then mapped into discrete labels as shown in Table 4.

$$H_f(C) := -\sum_{c_{ik}=C} \frac{n_k}{Z_f} log \frac{n_k}{Z_f} \tag{2}$$

$$H_b(C) := -\sum_{c_{jk}=C} \frac{n_k}{Z_b} log \frac{n_k}{Z_b} \tag{3}$$

where $Z_f = \sum_{c_{ik}=C} n_k, Z_b = \sum_{c_{jk}=C} n_k$ are the normalization factor, n_k denotes how many times the two consecutive characters c_{ik} and c_{jk} appear in corpus.

Table 4: Mapping from Conditional entropies to discrete labels

conditional entropies	labels
[0 ,1.0)	0
[1.0 ,2.0)	1
[2.0 ,3.5)	2
[3.5 ,5.0)	4
[5.0 ,7.0)	5
[7.0 ,$+\infty$)	6

Location Feature (LF) Location feature indicates position information of characters in words. Some characters are often used alone, and others may often be used as prefixes or suffixes of words. Yan [8] proved that the position feature has a positive effect on the improvement of the performance on word segmentation models. The way we introduce location feature is same as Yan [8].

Accessor Variety Feature (AV) AV is used to measure the independence of a string by checking the variety of its left and right neighbors. How we use AV is similar to Wu [10], In their model, only the first character in words is used. It does not make good use of the boundary information at the end of a substring. Hence, we designed a new set of templates as shown in Table 5. For character C, we use both the AV value of the substring ending with C and the AV value of the substring starting with C in templates. Considering the length of the substring is less than 6, we use the marker CAV1 represents AV of a substring whose length equals to 1, and so on.

Table 5: The templates of AV

length	Templates
1 char	$CAV1_{-1}$,$CAV1_0$,$CAV1_1$,$CAV1_{-1}CAV1_0$,$CAV1_0CAV1_1$,$CAV1_{-1}CAV1_1$
2 char	$CAV2_{-2}$,$CAV2_{-1}$,$CAV2_0$,$CAV2_1$,$CAV2_2$,$CAV2_{-2}CAV2_{-1}$, $CAV2_{-1}CAV2_0$,$CAV2_0CAV2_1$,$CAV2_{-1}CAV2_1$,$CAV2_1CAV2_2$
3 char	$CAV3_{-3}$,$CAV3_{-2}$,$CAV3_{-1}$,$CAV3_0$,$CAV3_1$,$CAV3_2$, $CAV3_{-3}CAV3_{-2}$, $CAV3_{-2}CAV3_{-1}$,$CAV3_{-1}CAV3_0$, $CAV3_0CAV3_1$, $CAV3_{-1}CAV3_1$,$CAV3_1CAV3_2$
4 char	$CAV4_{-4}$,$CAV4_{-3}$,$CAV4_{-2}$,$CAV4_{-1}$,$CAV4_0$,$CAV4_1$,$CAV4_2$, $CAV4_{-4}CAV4_{-3}$,$CAV4_{-3}CAV4_{-2}$,$CAV4_{-2}CAV4_{-1}$,$CAV4_{-1}CAV4_0$, $CAV4_0CAV4_1$,$CAV4_{-1}CAV4_1$,$CAV4_1CAV4_2$
5 char	$CAV5_{-5}$,$CAV5_{-4}$,$CAV5_{-3}$,$CAV5_{-2}$,$CAV5_{-1}$,$CAV5_0$,$CAV5_1$,$CAV5_2$, $CAV5_{-5}CAV5_{-4}$,$CAV5_{-4}CAV5_{-3}$,$CAV5_{-2}CAV5_{-1}$, $CAV5_{-1}CAV5_0$, $CAV5_0CAV5_1$,$CAV5_{-1}CAV5_1$,$CAV5_1CAV5_2$

Character Vector Feature (CVF) Distributed representation of words (characters) in vector space has been shown to be able to capture semantic information

which is also important for word segmentation. Two steps are taken for obtaining character vector feature. Firstly, each character is represented as a dense vector and trained by word2vec. Then we use K-means clustering to obtain its category information. Training for distributed representation usually needs a large-scale unsupervised data, so we only use the feature in the open track.

3 Experiment

There are three different tracks in this Shared Task. In closed and semi-open track, we only use the training data provided by organizer. MSRA and PKU data on Bakeoff-2005 [11] are joined in open track beside the data provided by organizer. Extra 1G Sina Weibo data is also used in open track, We combine those data with the training data to compute CEFs and AVs. CRF++[1] is used to implement our models.

3.1 Experimental Results

In this section, the official results of our models in different tracks are listed firstly. The models are evaluated by both standard measures [13] and weighted measures [2]. Since we have no official data of segmentation difficulty d_i for each word at that time, only standard measures (P, R, F1) are used to evaluate the efficiency of different features in our models latterly. Finally, we also present the results with weighted measures (p_b, r_b, f_b) when they were released by organizer.

The official results in different tracks are shown in Table 6 and Table 7, and byu-1 is our model.

Table 6: The official results with standard measures of our model on different tracks

Track	NickName (rank)	standard measures		
		P	R	F1
closed	panda-1 (1st)	0.939386	0.948021	0.947764
	bju (2st)	0.942164	0.953131	0.947616
	byu-1 (3st)	**0.943619**	**0.951535**	**0.94756**
semi-open	**byu-1 (1st)**	**0.947689**	**0.956219**	**0.951935**
	byu-2 (2st)	0.948134	0.955356	0.951732
	dlu-2 (3st)	0.946183	0.954941	0.950542
open	scu (1st)	0.950465	0.957007	0.953725
	bju (2st)	0.945906	0.955367	0.950613
	jj (3st)	0.935963	0.94659	0.941246
	byu-1 (4st)	**0.919183**	**0.914173**	**0.916671**

[1] http://crfpp.googlecode.com/svn/trunk/doc/index.html.

Table 7: The official results with weighted measures of our model on different tracks

Track	NickName (rank)	weighted measures		
		p_b	r_b	f_b
closed	byu-2 (1st)	0.792917	0.816246	0.804412
	bju (2st)	0.781889	0.818178	0.799622
	byu-1 (3st)	**0.7834**	**0.813453**	**0.798144**
semi-open	byu-2 (1st)	0.816755	0.843353	0.829841
	byu-1 (2st)	**0.804658**	**0.835274**	**0.81968**
	dlu-2 (3st)	0.793714	0.824671	0.808897
open	**byu-1 (1st)**	**0.812359**	**0.832221**	**0.82217**
	scu (2st)	0.803946	0.830007	0.816769
	bju (3st)	0.788601	0.821837	0.804876

For the closed track, We obtained 3nd place by two measures, with weighted measures, there is only 0.006 difference with the best result. On the semi-open track our model achieves the best F1 (0.951935), and 2^{th} place in f_b. Furthermore the f_b increased by 0.02 compared to the result on closed track. For the open track, our model achieves 4^{th} place by standard measures but the best with weighted measures.

We also show a series of experimental results on semi-open track with different features in Table 8. It could be seen that the performance is greatly improved after adding AV on the baseline with basic features. We tried two different set of templates for AV. One is from Wu [10], another is ours. It could be also seen that our templates for AV are more effective than that in Wu [10].

Table 8: Effect of features with standard measures of our model on semi-open tracks

Model	P	R	F1	P_{OOV}
Basic features (Baseline)	0.936	0.943	0.939	0.705
Basic features + AV (Wu [10])	0.943	0.953	0.948	0.702
Basic features + AV (Our)	0.945	0.954	0.949	0.708
+ LF, CEF	0.946	0.955	0.950	0.711

Finally, with all of the features, the model gets an F1 of 0.95. We can also see that the precision of OOV recognition is also improved from 0.705 to 0.711 along with different features added at the same time.

For the open track, due to the CVF has the same effect on expressing character type compared to CTF, we do some comparative experiments on the open track shown in table 9, and from the results we can see CVF is better than CTF.

The result also shows that the model with post processing gets F1 of 0.917 and gains improvement of 0.018. The post processing (PP) includes two aspects. One is non-Chinese character errors in micro-blog, like URL, E-mail address, decimals, percentages. Some regular expressions are designed for dealing with

them. The other is the Segmentation errors of Chinese character, an idiom dictionary[2] including about 23000 idioms is used in the open track, and atomic words (words without segmentation ambiguity) generated from training dataset are also used in our task.

Table 9: Effect of features with standard measures of our model on open tracks

Model	P	R	F1
Baseline[3]	0.891	0.897	0.894
Baseline+CTF	0.894	0.899	0.896
Baseline+CVF	0.896	0.901	0.899
Baseline+CVF+PP	0.919	0.914	0.917

Table 10: The results with weighted measures of our model on semi-open tracks

Model	p_b	r_b	f_b
Basic features (Baseline)	0.726	0.753	0.739
Basic features + AV (Wu [10])	0.791	0.827	0.809
Basic features + AV (Our)	0.799	0.833	0.815
+ LF, CEF	0.795	0.830	0.812
+ PP (We submit)	0.805	0.835	0.820
Basic features + AV (Our) + PP	0.807	0.838	0.822

There are many spaces remained to be improved in our system. After the weighted measures were released by organizer, we do some experiments on semi-open track with weighted measures and the results are listed on Table 10, from the table we can see that the performance declined after LF and CEF are added, which is a little different with Table 8. And the result is also better than the result we submitted comparing the last two lines, the reason for this phenomenon is that maybe there have the difference between the two evaluation measures. In this task, we select features based on the standard measures instead of weighted measures, so there may exist a better combination of features.

4 Conslusion

This paper describes our system for Chinese micro-blog segmentation. We mainly explore some features to improve the performance of the model. There are still many spaces remained to be improved in our system. In future, we can improve our methods from several aspects including exploring new features and new models special for micro-blog.

[2] https://github.com/sunflowerlyb/idiom
[3] Features including: CF, RF, AV, LF, CEF

Acknowledgments This work was partially supported by Natural Science Foundation of China (No.61273365), discipline building plan in 111 base (No.B08004) and Engineering Research Center of Information Networks of MOE, and the Co-construction Program with the Beijing Municipal Commission of Education.

References

1. Qiu, X., Qian, P., & Shi, Z., Wu, S., (2016). Overview of the NLPCC 2016 Shared Task: Chinese Word Segmentation for Micro-blog Texts.
2. Qian P, Qiu X, Huang X. A New Psychometric-inspired Evaluation Metric for Chinese Word Segmentation[C]// Meeting of the Association for Computational Linguistics. 2016.
3. Sebastiani F. Machine learning in automated text categorization[J]. Acm Computing Surveys, 2002, 34(1):1-47.
4. Jin K L, Ng H T, Guo W. A maximum entropy approach to Chinese word segmentation[J]. Proceedings of the Fourth Sighan Workshop on Chinese Language Processing, 2005.
5. Peng F, Feng F, Mccallum A. Chinese segmentation and new word detection using conditional random fields[J]. Proceedings of Coling, 2004:562–568.
6. Chen X, Qiu X, Zhu C, et al. Long Short-Term Memory Neural Networks for Chinese Word Segmentation[C]// Conference on Empirical Methods in Natural Language Processing. 2015.
7. Zhao H, Li M, Lu B L, et al. Effective Tag Set Selection in Chinese Word Segmentation via Conditional Random Field Modeling[C]// the 20th Pacific Asia Conference on Language,Information and Computation. 2006:87–94.
8. Jun Yan. Research and Application of Chinese Word Segmentation Based On Conditional Random Fields[D](in Chinese). 2009.
9. Gao Q, Vogel S. A Multi-layer Chinese Word Segmentation System Optimized for Out-of-domain Tasks[C]// 2010.
10. Wu, Guohua, et al. "Leveraging Rich Linguistic Features for Cross-domain Chinese Segmentation." Cips-Sighan Joint Conference on Chinese Language Processing 2014.
11. Emerson T. The second international chinese word segmentation bakeoff[C]//Proceedings of the fourth SIGHAN workshop on Chinese language Processing. 2005, 133.
12. Zhao H, Liu Q. The CIPS-SIGHAN CLP 2010 Chinese word segmentation bakeoff[C]//Proceedings of the First CPS-SIGHAN Joint Conference on Chinese Language Processing. 2010: 199-209.
13. Goutte C, Gaussier E. A probabilistic interpretation of precision, recall and F-score, with implication for evaluation[C]//European Conference on Information Retrieval. Springer Berlin Heidelberg, 2005: 345-359.