Micro blogs Oriented Word Segmentation System

Yijia Liu†, Meishan Zhang†, Wanxiang Che†, Ting Liu†, Yihe Deng‡
† Research Center for Social Computing and Information Retrieval
Harbin Institute of Technology, China
{yjliu,mszhang,car,tliu}@ir.hit.edu.cn
‡ School attached to Huazhong University of Science and Technology
Wuhan, 430074
brooklet60@gmail.com

Abstract

We present a Chinese word segmentation system submitted to the first task on CLP 2012 back-offs. Our segmenter is built using a conditional random field sequence model. We set the combination of a few annotated micro blogs and People Daily corpus as the training data. We encode special words detected by rules and information extracted from unlabeled data into features. These features are used to improve our model's performance. We also derive a micro blog specified lexicon from auto-analyzed data and use lexicon related features to assist the model. When testing on the sample data of this task, these features result in 1.8% improvement over the baseline model. Finally, our model achieves F-score of 94.07% on the bake-off's test set.

1 Introduction

Chinese word segmentation is the initial step of many NLP tasks, includes information retrieve, dependency parsing and semantic role labeling. Previous studies focus on word segmentation problem on standard data set, of which the training and testing data are drawn from same domain. However it's not always true when it comes to micro blogs. As a new source of information, micro blogs produce rich vocabulary ranging over many topics and changing with the times. Words like "给力" never appear in traditional data set, but occur frequently in micro blogs. At the same time, owing to the informal nature of micro blog, new type of words, such as URL, smiley and even the misspelled words, also make it very different from traditional task.

According to empirical analysis, one challenge of word segmentation on micro blogs is the sparsity issue resulting from lack of micro blog specified data. Current systems trained on standard data set perform poorly on micro blogs, because of domain mismatch. However, building a micro blogs specific word segmenter in standard supervised manner requires a lot of annotated data. Mannually creating them is a tedious and time-consuming work. Semi-supervised approaches, which make use of large scale unlabeled data is a promising solution to this issue. It enhances the segmenter with micro blog information and thus reduces sparsity in labeled training data. Recent studies have adopted semi-supervised approaches in word segmentation system(Wang et al., 2011; Sun and Xu, 2011), and improvement over the traditional supervised approach is observed.

Another challenge is the special word's detection. Due to the character of micro blogs, there are plentiful special words, such as hash tag, username, URL. Here is an example of micro blog entry: "[音乐] #我正在听# @MCHOTDOG熱狗《差不多先生》http://t.cn/h0VJQ （分享自@微博音乐盒）/ [music] #I'm listening# @MCHOTDOC *Mr. Ordinary* http://t.cn/h0VJQ (share from @weibomusicbox)". Words surrounded by "#" are hash tag, usually indicating the topics of the micro blog. "@MCHOTDOC熱狗" represent user names, and "http://t.cn/h0VJQ" is a shortened URL link. It's usually difficult for a word segmentation model to learn these changeable words from the training data. However, some certain type of special word can be detected by some rules easily and unambiguously. In this paper, we introduce some regular expressions to match special words in micro blog. The matching results, along with information extracted from un-

Proceedings of the Second CIPS-SIGHAN Joint Conference on Chinese Language Processing, pages 85–89,
Tianjin, China, 20-21 DEC. 2012

labeled data, are integrated into a CRF sequence model to learn a robust and high performance segmenter. We also derive a lexicon from auto-analyzed micro blog data and enhance our model with the lexicon information.

The reminder of this paper is organized as follows. Section 2 describes the details of our system. Section 3 presents experimental results and empirical analysis. Section 4 concludes this paper.

2 System Architecture

In this section, we describe the details of our system. We use some regular expressions to detect special words in micro blog. The detected word boundary of *URL*, *English word* and *special punctuation*, along with other information from unlabeled data, are integrated into a CRF sequence model as features. We build our first segmenter with information mentioned above and use this segmenter to parse large scale unlabeled data. After that, we extract a lexicon from auto-analyzed data and retrained the CRF model with information provided by the lexicon. The architecture of our system is illustrated in Figure 1.

2.1 Model and Basic Features

We employ a character-based sequence labeling model for word segmentation, which assign labels to the characters indicating whether a character is the beginning(B), inside(M), end of a word(E) or a unit-length word(S). A linear chain CRFs is used to learn model from annotated data. When considering the candidate character token c_i, the basic types of features of our model are listed below.

- character unigram: c_s $(i-2 \leq s \leq i+2)$
- character bigram: $c_s c_{s+1}$ $(i-2 \leq s \leq i+1)$, $c_s c_{s+2}$ $(i-2 \leq s \leq i)$
- character trigram: $c_{s-1} c_s c_{s+1}$ $(s=i)$
- repetition of characters: is c_s equals c_{s+1} $(i-1 \leq s \leq i)$, is c_s equals c_{s+2} $(i-2 \leq s \leq i)$
- character type: is c_i an *alphabet*, *digit*, *punctuation* or *others*

2.2 Rule Detection Features

We introduce regular expressions to detect three kinds of special words in micro blog, *URL*, *English word* and *Irregular suspension*. These three type of words are demonstrate as below.

- URL: "来 看 华 硕 新 版U36首 发 评 测 吧 ！ http://t.cn/aBPi3D / Come and see the reviews of newly released ASUS U36! http://t.cn/aBPi3D"
- English word: "分享Colbie Caillat 的歌曲/ Share Colbie Caillat's song"
- Irregular suspension: "非常的期待..... / I'm expecting ....."

We encode word boundary detected by the regular expressions into a new type of preprocessing features. If the candidate character token c_i, the following features about URL is extracted.

- beginning of a URL: $URL(c_i) = B$
- inside of a URL: $URL(c_i) = M$
- end of a URL: $URL(c_i) = E$

Features of *English word* and *irregular suspension* can be represented in same manner.

We expect that CRF model learns from these matching results and this information assists the CRF model to detect special words and words surrounding them.

2.3 Semi-supervised Features

Information of unlabeled data can be easily computed and benefit the word segmentation model. When integrated into machine learning framework, it will help reduce sparsity issue caused by the out of vocabulary words.

2.3.1 Mutual Information

In probability theory, mutual information measures the mutual dependency of two random variables. Empirical study shows that observation of high mutual information between two characters may indicates real association of these two characters in a word, while low mutual information usually means they belongs to different words.

In this paper, we follow Sun and Xu (2011)'s definition of mutual information. For a character bigram $c_i c_{i+1}$, their mutual information is computed as follow:

$$MI(c_i c_{i+1}) = \log \frac{p(c_i c_{i+1})}{p(c_i)p(c_{i+1})}$$

For each character c_i, $MI(c_i c_{i+1})$ and $MI(c_{i-1} c_i)$ are computed and rounded down to integer. We incorporate these values into our model as a type of features.

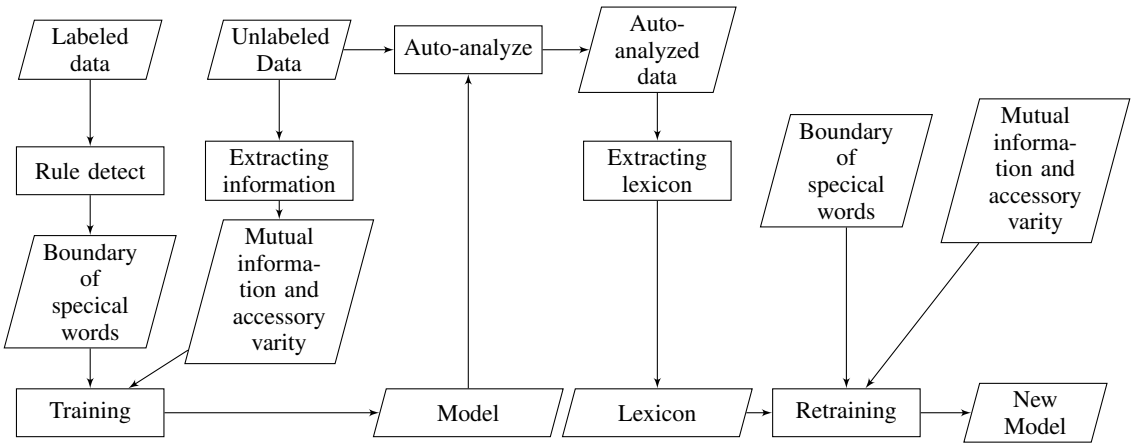

Figure 1: System architecture

2.3.2 Accessory Variety

Another empirical study of word segmentation boundary is that if some n-gram appears in many different environments, it's more likely that this n-gram be a real word. Sun and Xu (2011) introduce a criterion *Accessory Variety* to evaluate how independently a n-gram is used. In this paper, we follow this study and incorporate the following features $L^l_{AV}(c_{[i:i+l-1]})$, $L^l_{AV}(c_{[i+1:i+l]})$, $R^l_{AV}(c_{[i-l+1:i]})$, $R^l_{AV}(c_{[i-l:i-1]})$ $(l = 2, 3, 4)$ into our model. Here $L^l_{AV}(c_{[s:e]})$ and $R^l_{AV}(c_{[s:e]})$ means accessor variety of strings with length l, $c_{[s:e]}$ means the character sequence starts from c_s and ends with c_e.

2.4 Extracting Lexicon

Study has shown that CRF model can benefit from lexicon features(Zhang et al., 2010). Micro blog specified lexicon provides a clue for detecting words in unfamiliar context. In this paper, we try to extract a micro blog specified lexicon from auto-analyzed data to improve our model's performance.

Firstly, we train a CRF model with features described in 2.2 and 2.3. We use this model to parse large scale unlabeled data, and a list of word is obtained. Intuitively, high frequency word in the auto-analyzed results is more likely to be real word. Therefore, we collect words that never occur in the training data and rank them in order of frequency. A lexicon of words whose frequency is higher than a threshold is extracted. In this paper, top 80% most frequent words is extracted. We drop the tokens with more than 5 characters, and then build the lexicon.

After the lexicon D is built, we encode the information of lexicon into a type of features. We follow Zhang et al. (2010)'s work on utilization of lexicon When considering c_i, the lexicon feature we extract is shown below:

- $match\_prefix(c_i, D)$ the length of longest word in lexicon D which starts with c_i
- $match\_mid(c_i, D)$ the length of longest word in lexicon D which contains with c_i
- $match\_suffix(c_i, D)$ the length of longest word in lexicon D which ends with c_i

3 Experiments

3.1 Data Preparation and Setting

We crawl some micro blog from September 1st, 2011 to September 5nd, 2011, and drop the entries which not contains simplified Chinese characters. We got 1 million entries and use them as unlabeled data. From these micro blog entries, we randomly sampled 1,442 entries and manually annotated their segmentation. This set of corpus is use as one part of the labeled data. There are 23.3 words each entry in the annotated micro blogs on average. At the same time, 183,630 lines of sentences from People daily is also used as labeled data. All of the character in training and testing data is convert from single-byte character to double-byte character.

We use a toolkit - CRFSuite(Okazaki, 2007) to learning the sequence labeling model for segmentation. L-BFGS algorithm is set to solve the optimization problem.

We conclude our experiments result on the sample data of the bake-off task. There are 503 entries in the test data set, with 38.9 words each entry. Recall(R), precision(P) and F_1 is used as evaluation metrics of system performance. We also report the recall of out of vocabulary(OOV) words(R_{oov}).

3.2 Effect of Annotated Micro blog

In this set experiments, we test performance of standard supervised learning on different training data. As mentioned above, we have a large set of annotated corpus on newswire and a small set of micro blogs. We expected that a combination of these two corpus will help promote the performance.

We extract basic features from this two data and trained two CRF model BL^{pd} and BL^{mb}. Then we combine two data and trained another CRF model BL^{comb}. Performance of these three models is shown is Table 1.

Model	P	R	F
BL^{pd}	0.8820	0.8694	0.8757
BL^{mb}	0.8903	0.8925	0.8914
BL^{comb}	0.9161	0.9098	0.9130

Table 1: Effect of different annotated corpus

In previous study, the state-of-the-art word segmentation system can achieve F-score of about 97%(Che et al., 2010) when tested in-domain data. However, Table 1 shows that when applied to micro blogs, traditional word segmentation system's performance drops severely.

Experiment result also shows that, a small set of annotated micro blog corpus can achieve better performance than the traditional newswire corpus. And the model trained with combination of two corpus out performance the others. In the following section, all of our models are built on the combination of these two corpus.

3.3 Effect of Rule Detection Features

Table 2 compares the baseline model with model that integrates rule detection features.

Model	P	R	F	R_{oov}
BL	0.9161	0.9098	0.9130	0.5763
+PRE	0.9216	0.9178	0.9197	0.6715

Table 2: Effect of preprocessing

We can see that rule detection features improve the model's performance, especially the recall of OOV. To give a farther analysis of rule detection features' effect, we categorized words in test set into four sort: *URL*, *English word*, *Punctuation*, *Others* and evaluate the recall of certain type of word. Table 3 shows the experiment result.

Model	R_{URL}	R_{Punc}	R_{Eng}	R_{Others}
BL	0.8940	0.9857	0.6018	0.8997
+PRE	0.9536	0.9862	0.9227	0.9040

Table 3: Recall of preprocessing on four sort of words

The experiment result shows that rule detection features improves the recall of special word type, especially the English words occur in micro blog. With more accurate detection of sepecial words, accuracy on ordinary words is also improved.

3.4 Effect of Semi-supervised Features

Table 4 summarizes the experiment result on different combination of semi-supervised features.

Model	P	R	F	R_{oov}
BL+PRE	0.9216	0.9178	0.9197	0.6715
+MI	0.9282	0.9220	0.9251	0.7046
+AV	0.9309	0.9231	0.9270	0.7250
+MI+AV	0.9304	0.9231	0.9268	0.7123

Table 4: Effect of semi-supervised features

It can be seen that two types of semi-supervised features both result in improvement on performance. However, when two types of feature combined, the performance drops slightly. Empirically, we consider that the effect of these two type features overlaps due to they share some common property.

3.5 Effect of Lexicon

We also compare our model integrating lexicon features and without lexicon features. The results are shown in Table 5.

Model	P	R	F	R_{oov}
BL+PRE+MI+AV	0.9304	0.9231	0.9268	0.7123
+Lexicon	0.9352	0.9275	0.9314	0.7337

Table 5: Effect of lexicon features

As expected, lexicon features result in improvement over performance.

3.6 Final System

Our final system is set as the configuration of "BL+PRE+MI+AV+Lexicon". Our experimental

results show that our final system achieves an F-score of 93.14% and an improvement of 1.8% comparing to our baseline model. On the evaluation data of the bake-off, the F-score of our system is 94.07%.

4 Conclusion

In this paper, we describe our system of *Chinese Word Segmentation on MicroBlog Corpora*. We exploit a single model enhanced by preprocessing, semi-supervised and lexicon features. These features improve the model's performance. Our model achieve an F-score of 94.07% on the bake-off's test data.

Acknowledgments

This work was supported by National Natural Science Foundation of China (NSFC) via grant 61133012, the National "863" Major Projects via grant 2011AA01A207, and the National "863" Leading Technology Research Project via grant 2012AA011102.

References

W. Che, Z. Li, and T. Liu. 2010. Ltp: A chinese language technology platform. In *Proceedings of the 23rd International Conference on Computational Linguistics: Demonstrations*, pages 13–16. Association for Computational Linguistics.

Naoaki Okazaki. 2007. Crfsuite: a fast implementation of conditional random fields (crfs).

W. Sun and J. Xu. 2011. Enhancing chinese word segmentation using unlabeled data. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 970–979. Association for Computational Linguistics.

Y. Wang, Y.T. Jun' ichi Kazama, W. Chen, Y. Zhang, and K. Torisawa. 2011. Improving chinese word segmentation and pos tagging with semi-supervised methods using large auto-analyzed data. In *Proceedings of the Fifth International Joint Conference on Natural Language Processing (IJCNLP-2011)*.

M. Zhang, Z. Deng, W. Che, and T. Liu. 2010. Combining statistical model and dicitionary for domain adaptation of chinese word segmentation. *Journal of Chinese Information Processing*, 26(2):8–12.