Overview of the NLPCC-ICCPOL 2016 Shared Task: Chinese Word Segmentation for Micro-blog Texts

Xipeng Qiu, Peng Qian, Zhan Shi

School of Computer Science, Fudan University
825 Zhangheng Road, Shanghai, China
{xpqiu,pqian11,zshi16}@fudan.edu.cn

Abstract. In this paper, we give an overview for the shared task at the 5th CCF Conference on Natural Language Processing & Chinese Computing (NLPCC 2016): Chinese word segmentation for micro-blog texts. Different with the popular used newswire datasets, the dataset of this shared task consists of the relatively informal micro-texts. Besides, we also use a new psychometric-inspired evaluation metric for Chinese word segmentation, which addresses to balance the very skewed word distribution at different levels of difficulty. The data and evaluation codes can be downloaded from https://github.com/FudanNLP/NLPCC-WordSeg-Weibo.

1 Introduction

Word segmentation is a fundamental task for Chinese language processing. Benefiting from the developments of the machine learning techniques and the large scale shared corpora, Chinese word segmentation has achieved a great progress. The state-of-the-art method is to regard this task as sequence labeling problem. However, their performances are still not satisfying for the practical demands to analyze Chinese texts, especially for informal texts. The key reason is that most of annotated corpora are drawn from news texts. Therefore, the system trained on these corpora cannot work well with the informal or specific-domain texts. To address this, we introduce a new large corpus and a new evaluation metric [3]. We hope that our corpus and metric can provide a valuable testbed for Chinese word segmentation on informal texts.

In this shared task, we wish to investigate the performances of Chinese word segmentation for the micro-blog texts. Different with the former task in NLPCC 2015 [4], we just focus on word segmentation and introduce a new evaluation metric [3] this year.

2 Data

Different with the popular used newswire dataset, we use relatively informal texts from Sina Weibo[1]. The training and test data consist of micro-blogs from various topics, such as finance, sports, entertainment, and so on. Both the training and test files are UTF-8 encoded. To reduce the cost of data annotation, we use FudanNLP[2] [5] to obtain the

[1] http://weibo.com/
[2] https://github.com/FudanNLP/fnlp

initial segmentations. Then two annotators modify the errors in the initial segmentation. When two annotators disagree, a third annotator gives a final decision.

The information of dataset is shown in Table 1. The out-of-vocabulary (OOV) rate is slight higher than the other benchmark datasets.

Table 1: Statistical information of dataset.

DataSet	Sents	Words	Chars	Word Types	Char Types	OOV Rate
Train	20135	421166	688743	43331	4502	-
Develop	2052	43697	73246	11187	2879	6.82%
Test	8592	187877	315865	27804	3911	6.98%
Total	30779	652740	1077854	56155	4838	-

2.1 Background Data

Besides the training data, we also provide the background data, from which the training and test data are drawn. The purpose of providing the background data is to find the more sophisticated features by the unsupervised way.

3 Description of the Task

Word is the fundamental unit in natural language understanding. However, Chinese sentences consists of the continuous Chinese characters without natural delimiters. Therefore, Chinese word segmentation has become the first mission of Chinese natural language processing, which identifies the sequence of words in a sentence and marks the boundaries between words.

3.1 Tracks

Each participant will be allowed to submit the three runs for each subtask: **closed track** run, **semi-open track** run and **open track** run.

1. In the **closed** track, participants could only use information found in the provided training data. Information such as externally obtained word counts, part of speech information, or name lists was excluded.
2. In the **semi-open** track, participants could use the information extracted from the provided background data in addition to the provided training data. Information such as externally obtained word counts, part of speech information, or name lists was excluded.
3. In the **open** track, participants could use the information which should be public and be easily obtained. But it is not allowed to obtain the result by the manual labeling or crowdsourcing way.

4 Evaluations

4.1 Evaluation Metric

After the successive improvements, the standard metric is becoming hard to distinguish state-of-the-art word segmentation systems. In this shared task, we use a new psychometric-inspired evaluation metric for Chinese word segmentation, which addresses to balance the very skewed word distribution at different levels of difficulty. The performance on a real evaluation shows that the proposed metric gives more reasonable and distinguishable scores and correlates well with human judgement. The detailed information can be found in [3].

To show the difference between the standard and new evaluation measures, we report both metrics for each system.

4.2 Results

There are 19 submitted systems. The results on three tracks are shown in Table 2, 3 and 4 respectively.

Table 2: Performances on closed track.

	Standard Scores			Weighted Scores		
Systems	p	r	f_1	p_b	r_b	f_b
S1	94.13	94.69	94.41	**79.29**	81.62	**80.44**
S2	94.21	**95.31**	**94.76**	78.18	**81.81**	79.96
S3	**94.36**	95.15	94.75	78.34	81.34	79.81
S4	93.98	94.78	94.38	78.43	81.2	79.79
S5	93.93	94.8	94.37	76.24	79.32	77.75
S6	93.9	94.42	94.16	75.95	78.2	77.06
S7	93.82	94.6	94.21	75.08	77.91	76.47
S8	93.74	94.31	94.03	74.9	77.14	76
S9	92.89	93.65	93.27	71.25	73.92	72.56
S10	93.31	93.83	93.57	71.22	73.32	72.25
S11	93.52	94.14	93.83	70.12	72.55	71.31
S12	90.78	91.88	91.33	68.29	71.93	70.06
S13	87.93	89.82	88.86	61.05	66.06	63.46
S14	85.08	87.18	86.12	55.04	59.77	57.31
S15	66.39	73.6	69.81	50	63.84	56.08
S16	80.53	80.53	80.53	41.3	43.61	42.42
Average	90.16	91.42	90.8	69.04	72.72	70.8

4.3 Some Representative Systems

In this section, we give brief introductions to some representative system.

Table 3: Performances on semi-open track.

	Standard Scores			Weighted Scores		
Systems	p	r	f_1	p_b	r_b	f_b
S1	**94.81**	95.53	95.17	**81.67**	**84.33**	**82.98**
S3	94.76	**95.62**	**95.19**	80.46	83.52	81.96
S8	94.61	95.49	95.05	79.37	82.46	80.88
S7	94.56	95.35	94.95	78.98	81.8	80.36
S2	94.23	95.32	94.77	78.2	81.87	79.99
S9	90.49	91.76	91.12	68.3	72.24	70.21
S16	80.32	79.91	80.12	41.64	43.55	42.57
AVG	91.97	92.71	92.34	72.66	75.68	74.14

Table 4: Performances on open track.

	Standard Scores			Weighted Scores		
Systems	p	r	f_1	p_b	r_b	f_b
S3	91.91	91.41	91.66	**81.23**	**83.22**	**82.21**
S11	**95.04**	**95.7**	**95.37**	80.39	83	81.67
S2	94.59	95.53	95.06	78.86	82.18	80.48
S6	93.59	94.65	94.12	75.86	79.42	77.6
S9	90.78	91.88	91.33	68.29	71.93	70.06
AVG	93.18	93.83	93.51	76.93	79.95	78.41

– The **S1** system [7] uses Long Short-Term Memory (LSTM) for Chinese Weibo word segmentation. In order to infer the optimal tag path, a transition score matrix is used for jumping between tags in successive characters. By integrating unsupervised features, the performance is further improved.
– The **S3** system [2] uses sequence labeling for CWS with CRF model. It takes full advantages of both unsupervised features and supervised features to discover new words from unlabeled dataset. The new words recognition is significantly improved with those features.
features. Accessor variety (AV) [1] features are used to measure the possibility of whether a substring is a Chinese word. They report that the ability of OOV detection can be improved by integrating unsupervised global features extracted from the provided background data.
– The **S11** system [6] also treats word segmentation as a character-wise sequence labeling problem, and explores two directions to enhance the CRF-based baseline. First, a large-scale external lexicon is employed for constructing extra lexicon features in the model, which is proven to be extremely useful. Second, two heterogeneous datasets, i.e., Penn Chinese Treebank 7 (CTB7) and People Daily (PD) are used to help word segmentation on Weibo.

5 Analysis

The analyses of the participant systems are as follows.

1. The best system on semi-open track is better than that on closed track, which shows the large scale unlabeled data from the same domain are useful for Chinese word segmentation.
2. The neural network based model shows a distinct advantage. The **S1** system adopts LSTM to model the sequence and achieves the best results on both closed and semi-open tracks.
3. The new evaluation metric gives more distinguishable score than the standard metric.

6 Conclusion

After years of intensive researches, Chinese word segmentation has achieved a quite high precision. However, the performances of state-of-the-art systems are still relatively low for the informal texts, such as micro-blogs, forums. The NLPCC 2016 Shared Task on Chinese Word Segmentation for Micro-blog Texts focuses on the fundamental research in Chinese language processing. It is the first time to use the micro-texts to evaluate the performance of the state-of-the-art methods. Besides, we also wish to extend the scale of corpus and add more informal texts.

Acknowledgement

We are very grateful to the students from our lab for their efforts to annotate and check the data. We would also like to thank the participants for their valuable feedbacks and comments. This work was partially funded by National Natural Science Foundation of China (No. 61532011 and 61672162), the National High Technology Research and Development Program of China (No. 2015AA015408).

References

1. Feng, H., Chen, K., Deng, X., Zheng, W.: Accessor variety criteria for chinese word extraction. Computational Linguistics 30(1), 75–93 (2004)
2. Leng, Y., Liu, W., Wang, S., , Wang, X.: A feature-rich CRF segmenter for chinese micro-blog. In: Proceedings of The Fifth Conference on Natural Language Processing and Chinese Computing & The Twenty Fourth International Conference on Computer Processing of Oriental Languages (2016)
3. Qian, P., Qiu, X., Huang, X.: A new psychometric-inspired evaluation metric for chinese word segmentation. In: Proceedings of Annual Meeting of the Association for Computational Linguistics (2016), http://aclweb.org/anthology/P/P16/P16-1206.pdf
4. Qiu, X., Qian, P., Yin, L., Huang, X.: Overview of the NLPCC 2015 shared task: Chinese word segmentation and POS tagging for micro-blog texts. In: National CCF Conference on Natural Language Processing and Chinese Computing. pp. 541–549. Springer (2015), http://arxiv.org/abs/1505.07599
5. Qiu, X., Zhang, Q., Huang, X.: FudanNLP: A toolkit for Chinese natural language processing. In: Proceedings of Annual Meeting of the Association for Computational Linguistics (2013)

6. Xia, Q., Li, Z., Chao, J., Zhang, M.: Word segmentation on micro-blog texts with external lexicon and heterogeneous data. In: Proceedings of The Fifth Conference on Natural Language Processing and Chinese Computing & The Twenty Fourth International Conference on Computer Processing of Oriental Languages (2016)
7. Zhou, Q., Ma, L., Zheng, Z., Wang, Y., Wang, X.: Recurrent neural word segmentation with tag inference. In: Proceedings of The Fifth Conference on Natural Language Processing and Chinese Computing & The Twenty Fourth International Conference on Computer Processing of Oriental Languages (2016)