Overview of the NLPCC 2015 Shared Task: Chinese Word Segmentation and POS Tagging for Micro-blog Texts

Xipeng Qiu, Peng Qian, Liusong Yin, Shiyu Wu, Xuanjing Huang
School of Computer Science, Fudan University
825 Zhangheng Road, Shanghai, China
{xpqiu,pqian11,lsyin14,sywu13,xjhuang}@fudan.edu.cn

Abstract

In this paper, we give an overview for the shared task at the 4th CCF Conference on Natural Language Processing & Chinese Computing (NLPCC 2015): Chinese word segmentation and part-of-speech (POS) tagging for micro-blog texts. Different with the popular used newswire datasets, the dataset of this shared task consists of the relatively informal micro-texts. The shared task has two sub-tasks: (1) individual Chinese word segmentation and (2) joint Chinese word segmentation and POS Tagging. Each subtask has three tracks to distinguish the systems with different resources. We first introduce the dataset and task, then we characterize the different approaches of the participating systems, report the test results, and provide a overview analysis of these results. An online system is available for open registration and evaluation at `http://nlp.fudan.edu.cn/nlpcc2015`.

1 Introduction

Word segmentation and Part-of-Speech (POS) tagging are two fundamental tasks for Chinese language processing. In recent years, word segmentation and POS tagging have undergone great development. The popular method is to regard these two tasks as sequence labeling problem [7, 5], which can be handled with supervised learning algorithms such as Maximum Entropy (ME) [1], averaged perceptron [2], Conditional Random Fields (CRF)[3]. After years of intensive researches, Chinese word segmentation and POS tagging achieve a quite high precision. However, their performance is not so 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 out-of-domain texts.

In this shared task, we focus to evaluate the performances of word segmentation and POS tagging on relatively informal micro-texts.

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.

The information of dataset is shown in Table 1. The out-of-vocabulary (OOV) rate is slight higher than the other benchmark datasets. For example, the OOV rate is 5.58% in the popular division [9] of the Chinese Treebank (CTB 6.0) dataset [8], while the OOV rate of our dataset is 7.25%.

Table 1: Statistical information of dataset.

Dataset	Sents	Words	Chars	Word Types	Char Types	OOV Rate
Training	10,000	215,027	347,984	28,208	39,71	-
Test	5,000	106,327	171,652	18,696	3,538	7.25%
Total	15,000	322,410	520,555	35,277	4,243	-

There are total 35 POS tags in this dataset. A detailed list of POS tags is shown in Table 2.

[1]http://weibo.com/

Table 2: Statistical information of POS tags.

词性(POS)		En	Num
名词		NN	84,006
实体名	人名	PER	3,232
	机构名	ORG	2,578
	地名	LOC	9,701
	其他	NR	550
	邮件	EML	3
	型号名	MOD	34
	网址	URL	11
副词	疑问副词	ADQ	340
	副词	AD	26,155
形貌	形容词	JJ	9,477
	形谓词	VA	3,339
动词	动词	VV	51,294
	情态词	MV	3,700
	趋向动词	DV	781
	被动词	BEI	927
	把动词	BA	600
时间短语		NT	5,881

词性(POS)		Labels	Num
代词	人称代词	PNP	4,903
	疑问代词	PNQ	492
	指示代词	PNI	834
连词	并列连词	CC	2,725
	从属连词	CS	866
数量	数词	CD	10,764
	量词	M	7,917
	序数词	OD	1,219
助词	方位词	LC	4,725
	省略词	ETC	673
	语气词	SP	1,076
	限定词	DT	3,579
	叹词	IJ	20
	标点	PU	52,922
	结构助词	DSP	13,756
	介词	P	9,488
	时态词	AS	3,382

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 is to find the more sophisticated features by the unsupervised way.

3 Description of the Task

In this shared task, we wish to investigate the performances of Chinese word segmentation and POS tagging for the micro-blog texts.

3.1 Subtasks

This task focus the two fundamental problems of Chinese language processing: word segmentation and POS tagging, which can be divided into two subtasks:

1. **SEG** Chinese word segmentation
2. **S&T** Joint Chinese word segmentation and POS Tagging

3.2 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 Participants

Sixteen teams have registered for this task. Finally, there are 27 qualified submitted results from 10 teams. A summary of qualified participating teams are shown in Table 3.

Table 3: Summary of the participants.

	SEG			S&T		
	closed	open	semi-open	closed	open	semi-open
NJU	√	√	√			
BosonNLP	√	√		√	√	
CIST	√		√	√		√
XUPT	√			√		
CCNU	√	√				
ICT-NLP	√					
BJTU	√	√	√	√	√	√
SZU		√			√	
ZZU			√			
WHU				√		√

5 Results

5.1 Evaluation Metrics

The evaluation measure are reported are precision, recall, and an evenly-weighted F1.

5.2 Baseline Systems

Currently, the mainstream method of word segmentation is discriminative character-based sequence labeling. Each character is labeled as one of $\{$B, M, E, S$\}$ to indicate the segmentation. $\{$B, M, E$\}$ represent *Begin*, *Middle*, *End* of a multi-character segmentation respectively, and S represents a *Single* character segmentation.

For the joint word segmentation and POS tagging, the state-of-the-art method is also based on sequence learning with cross-labels, which can avoid the problem of error propagation and achieve higher performance on both subtasks[4]. Each label is the cross-product of a segmentation label and a tagging label, e.g. $\{$B-NN, I-NN, E-NN, S-NN, ...$\}$. The features are generated by position-based templates on character-level.

Sequence labeling is the task of assigning labels $\mathbf{y} = y_1, \ldots, y_n$ to an input sequence $\mathbf{x} = x_1, \ldots, x_n$. Given a sample $\mathbf{x}$, we define the feature $\Phi(\mathbf{x}, \mathbf{y})$. Thus, we can label $\mathbf{x}$ with a score function,

$$\hat{\mathbf{y}} = \arg\max_{\mathbf{y}} F(\mathbf{w}, \Phi(\mathbf{x}, \mathbf{y})), \tag{1}$$

where $\mathbf{w}$ is the parameter of function $F(\cdot)$.

For sequence labeling, the feature can be denoted as $\phi_k(y_i, y_{i-1}, \mathbf{x}, i)$, where i stands for the position in the sequence and k stands for the number of feature templates.

Here, we use two popular open source toolkits for sequence labeling task as the baseline systems: FNLP[2] [6] and CRF++[3]. Here, we use the default setting of CRF++ toolkit with the feature templates as shown in Table 4. The same feature templates are also used for FNLP.

5.3 Chinese word segmentation

In word segmentation task, the best F1 performances are 95.12, 95.52 and 96.65 for closed, semi-open and open tracks respectively. The best system outperforms the baseline systems on closed track. The best system on semi-open track is better

[2] `https://github.com/xpqiu/fnlp/`

[3] `http://taku910.github.io/crfpp/`

Table 4: Templates of CRF++ and FNLP.

unigram feature	c_{-2}, c_{-1}, c_0, c_{+1}, c_{+2}
bigram feature	$c_{-1} \circ c_0$, $c_0 \circ c_{+1}$
trigram feature	$c_{-2} \circ c_{-1} \circ c_0$, $c_{-1} \circ c_0 \circ c_{+1}$, $c_0 \circ c_{+1} \circ c_{+2}$

than that on closed track. Unsurprisingly, the performances boost greatly on open track.

Table 5: Performances of word segmentation.

Systems	Precision	Recall	F1	Track
CRF++	93.3	93.2	93.3	baseline, closed
FNLP	94.1	93.9	94.0	
NJU	95.14	95.09	95.12	closed
BosonNLP	95.03	95.03	95.03	
CIST	94.78	94.42	94.6	
XUPT	94.61	93.85	94.22	
CCNU	93.95	93.45	93.7	
ICT-NLP	93.96	92.91	93.43	
BJTU	89.49	93.55	91.48	
CIST	95.47	95.57	95.52	semi-open
NJU	95.3	95.31	95.3	
BJTU	90.91	94.46	92.65	
ZZU	85.36	85.25	85.31	
BosonNLP	96.56	96.75	96.65	open
NJU	96.03	96.15	96.09	
SZU	95.52	95.64	95.58	
CCNU	93.68	93.09	93.38	
BJTU	91.79	94.92	93.33	

5.4 Joint Chinese word segmentation and POS Tagging

In the joint word segmentation and POS tagging, the best performances are 88.93, 88.69 and 91.55 for closed, semi-open and open tracks respectively.

Table 6: Performances of joint word segmentation and POS tagging.

Systems	Precision	Recall	F1	Track
BosonNLP	88.91	88.95	88.93	closed
XUPT	88.54	87.83	88.19	
BJTU	88.28	87.67	87.97	
CIST	88.09	87.76	87.92	
BJTU	80.64	85.1	82.81	
CIST	88.64	88.73	88.69	semi-open
WHU	88.59	87.96	88.27	
BJTU	81.76	85.82	83.74	
BosonNLP	91.42	91.68	91.55	open
SZU	88.93	89.05	88.99	
BJTU	79.85	83.51	81.64	

6 Analysis

7 Conclusion

After years of intensive researches, Chinese word segmentation and POS tagging have achieved a quite high precision. However, the performances of the state-of-the-art systems are still relatively low for the informal texts, such as micro-blogs, forums. The NLPCC 2015 Shared Task on Chinese Word Segmentation and POS Tagging 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

In future work, we hope to run an online evaluation system to accept open registration and submission. Currently, a simple system is available at `http://nlp.fudan.edu.cn/nlpcc2015`. The system also gives the leaderboards for the up-to-date results under the different tasks and tracks. 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.

References

[1] A.L. Berger, V.J. Della Pietra, and S.A. Della Pietra. A maximum entropy approach to natural language processing. *Computational Linguistics*, 22(1):39–71, 1996.

[2] Michael Collins. Discriminative training methods for hidden markov models: Theory and experiments with perceptron algorithms. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing*, 2002.

[3] John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, 2001.

[4] H.T. Ng and J.K. Low. Chinese part-of-speech tagging: one-at-a-time or all-at-once? word-based or character-based. In *Proceedings of EMNLP*, volume 4, 2004.

[5] F. Peng, F. Feng, and A. McCallum. Chinese segmentation and new word detection using conditional random fields. *Proceedings of the 20th international conference on Computational Linguistics*, 2004.

[6] Xipeng Qiu, Qi Zhang, and Xuanjing Huang. FudanNLP: A toolkit for Chinese natural language processing. In *Proceedings of Annual Meeting of the Association for Computational Linguistics*, 2013.

[7] N. Xue. Chinese word segmentation as character tagging. *Computational Linguistics and Chinese Language Processing*, 8(1):29–48, 2003.

[8] Naiwen Xue, Fei Xia, Fu-Dong Chiou, and Martha Palmer. The Penn Chinese TreeBank: Phrase structure annotation of a large corpus. *Natural language engineering*, 11(2):207–238, 2005.

[9] Yaqin Yang and Nianwen Xue. Chinese comma disambiguation for discourse analysis. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 786–794. Association for Computational Linguistics, 2012.