Shared Task in NLPCC 2015: Chinese Word Segmentation and POS Tagging for Weibo Text

April 11, 2015

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, which can be handled with supervised learning algorithms such as Conditional Random Fields (CRF). However, the performances of the state-of-the-art systems are still relatively low for the informal texts, such as micro-blogs, forums. In this shared task, we wish to investigate the performances of Chinese word segmentation and POS tagging for the micro-blog texts.

2 Description of the Task

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

1. Chinese word segmentation
2. Joint Chinese word segmentation and POS Tagging

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

Dataset	Sents	Words	Chars	Word Types	Char Types	OOV Rate
Training	10,000	215,567	348,551	28,355	39,73	-
Test	5,000	106,843	172,342	18,785	3,540	9.75%
Total	15,000	322,410	520,555	35,277	4,243	-

Table 1: Statistical information of dataset.

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.

A website will be provided to evaluate the results.

3 Data

Different with the popular used news 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.

The data are collected from Sina Weibo. Both the training and test files are UTF-8 encoded. The information of dataset is shown in Table 1.

There are total 36 POS tags in this dataset. A detailed list of POS tags is shown in Table 2.

3.1 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.

4 Evaluation Metrics

We use the standard SIGHAN bake-off scoring program to calculate precision, recall, F1-score and out-of-vocabulary (OOV) word recall.

A website will be available for automatical online evaluation when test data are released.

[1] http://weibo.com/

词性 (POS)		Labels	Occurrences
名词		NN	84,282
实体名	人名	PER	3,192
	机构名	ORG	2,254
	地名	LOC	9,644
	其他	NR	597
	邮件	EML	3
	型号名	MOD	24
	专有名	SN	92
	网址	URL	11
副词	疑问副词	ADQ	316
	副词	AD	26,440
形貌	形容词	JJ	9,699
	形谓词	VA	3,277
动词	动词	VV	51,344
	情态词	MV	3,640
	趋向动词	DV	806
	被动词	BEI	923
	把动词	BA	600
代词	人称代词	PNP	4,868
	疑问代词	PNQ	490
	指示代词	PNI	844
连词	并列连词	CC	2,733
	从属连词	CS	869
数量	数词	CD	10,829
	量词	M	7,926
	序数词	OD	1,217
时间短语		NT	5,831
助词	方位词	LC	4,788
	省略词	ETC	673
	语气词	SP	1,068
	限定词	DT	3,606
	叹词	IJ	20
	标点	PU	52,865
	结构助词	DSP	13,761
	介词	P	9,489
	时态词	AS	3,389

Table 2: Statistical information of POS tags.