NLPCC-ICCPOL 2016 Shared Task Guideline:
Chinese Word Segmentation for Weibo Text

1 Introduction

Word segmentation is a fundamental task for Chinese language processing. In recent years, word segmentation has 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 for the micro-blog texts.

2 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.

Different with the popular used news dataset, we use more informal texts from Sina Weibo. The training and test data consist of micro-blogs from various topics, such as finance, sports, entertainment, and so on.

Each participant will be allowed to submit the three runs: **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.

3 Data

The data are collected from Sina Weibo. Both the training and test files are UTF-8 encoded. 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.

4 Evaluation Metric

Different with the standard precision, recall, F1-score, we will provide a new measure metric this year. The detailed metric will be available when we release the training and test datasets.

5 Contact Information

For any questions about this shared task, please contact:

Xipeng Qiu

School of Computer Science, Fudan University

Email: xpqiu@fudan.edu.cn