Chinese Tweets Segmentation based on Morphemes

Chaoyue Wang
Heilongjiang University
Harbin, China.
chaoyue.wang@yahoo.cn

Guohong Fu
Heilongjiang University
Harbin, China.
ghfu@hlju.edu.cn

Abstract

Chinese tweets segmentation is a critical problem in natural language processing area. While segmentation of in-vocabulary words is well studied to date, few research findings are yet available concerning the prediction of new words on twitter. In this paper, we attempt to exploit multiple features for segmenting tweets in real text. To this end, we first take morpheme as the basic component units of Chinese words and thus investigate the relationship between Chinese new words and their internal morphological structures. Then, we explore both word internal cues and word external contextual features, and combine them for segmentation of Chinese new words using conditional random field. Our experimental results show that the incorporation of multiple features, especially the word-internal morphological features is of great value to Chinese tweets segmentation.

1 Introduction

Chinese word segmentation is one of the important steps in natural language processing. Essentially, segmentation is trying to determine the boundary of the word. As a fundamental natural language analysis task, word segmentation plays a key role in many natural language processing applications.

Different from the traditional word segmentation, many new words exist in the segmentation on twitter. Traditional methods can't deal with this problem well, especially the dictionary based method. In this paper, we use statistical method to solve this problem.

In previous study, most researchers used word as the basic unit; however, this method is fatigue on addressing the new words detection. To address this problem, in this paper, we use morpheme as the basic unit under the Conditional Random Filed (CRF). Fu et al. proved that morphemes were informative for unknown words processing.

2 Approach

In this paper, we take word segmentation as sequence labeling. Given an input sequence of words, our approach for word segmentation consists of three main parts: First, a word decomposition module is employed to decompose unknown words within the input sentence into a sequence of morphemes. Then the extended BIO tagset is used to represent the position patterns of morphemes within words. Finally, CRF is used to predict the corresponding label.

2.1 Chinese Morphemes

In the present study we consider two major types of morphemes, namely free morphemes and bound morphemes (viz. affixes). A free morpheme can stand by itself as a word, whereas an affix can show up if and only if being attached to other morphemes to form a word.

To explore word-internal clues for segmentation of Chinese new words, we employ the extended IOB tagset to represent the position patterns of Chinese morphemes in word formation. Table 1 presents the detailed definition of the extended IOB tags and the correspondence between IOB tags and morpheme types.

Tag	Definition
O	A morpheme as a word by itself
I	A morpheme inside a word
B	A word-initial morpheme
E	A word-final morpheme

Table1 The extended IOB tagset for the representation of component morphemes within Chinese word

Proceedings of the Second CIPS-SIGHAN Joint Conference on Chinese Language Processing, pages 106–107,
Tianjin, China, 20-21 DEC. 2012

2.2 Word decomposition

Word decomposition is the process of decomposing a word to a sequence of morphemes associated with their IOB tags defined in Table 1. For example, the word "不安全感"(the sense of insecurity) should be decomposed as "不/O 安/B 全/E 感/O".

2.3 Features

Feature selection plays a critical in CRF. In the present study, we consider two main groups of features for Chinese word segmentation, namely contextual features around words and word-formation features within words. We choose the part of speech (POS) of the morpheme as the internal feature; the table2 shows our feature template.

```
# Unigram
    U00:%x[-1,0]
    U01:%x[0,0]
    U02:%x[1,0]
    U03:%x[-1,1]
    U04:%x[0,1]
    U05:%x[1,1]
    U06:%x[-1,0]/%x[0,0]
    U07:%x[0,0]/%x[1,0]
    U08:%x[-1,1]/%x[0,1]
    U09:%x[0,1]/%x[1,1]
    # Bigram
    B
```

Table 2 Feature template for morpheme-based CRFs

3 Experimental result

Table 3 shows the result. The 'Best' indicates the high score achieved in CLP2012 Micro-blog word segmentation subtask.

Results	Precision Rate	Recall Rate	F Score	Total Correct Sentences	Ratio of Correct Sentences
Our Result	0.8451	0.8437	0.8444	750	15.0%
Best	0.946	0.9496	0.9478	2244	44.88%

Table 3 Evaluation Results

4 Conclusions

In this paper, we have attempted to explore word internal morphological clues within Chinese words, and incorporate them with word-external contextual features for segmentation of Chinese words. Due to the lack of large scale corpus and deep morphological knowledge for Chinese, in the present study we only took into account surface morphological clues, namely the position patterns of morphemes in word formation. In future work we intend to explore systematically deep morphological knowledge.

References

Ruiqiang Zhang, Keiji Yasuda, Eiichiro Sumita. 2008. Improved statistical machine traslantion by multiple Chinese word segmentation. Proceedings of the 3rd workshop on statistical machine translation, 216-223.

S. Foo, H. Li. 2004. Chinese word segmentation and its effect on information retrieval. Information Processing and Management, 40(1): 161-190.

Guohong Fu, Kang-Kwong Luke. 2006. Chinese POdisambiguation and unknown word guessing with lexicalized HMMs. International Journal of Technology and Human Interaction, Vol.2, No.1, pages 39-50.

Guohong Fu, Chunyu Kit, Jonathan J. 2008. Webster. Chinese word segmentation as morpheme-based lexical chunking. Information Sciences, Vol.178, No.9, pages 2282-2296.