Chinese Word Segmentation Based On Direct Maximum Entropy Model

Wu-Guang Shi
Peking University, Beijing, 100871, China
shiwuguang@pku.edu.cn

Abstract

Chinese word segmentation is a fundamental and important issue in Chinese information processing. In order to find a unified approach for Chinese word segmentation, the author develop a Chinese lexical analyzer PCWS using direct maximum entropy model. The paper presents the general description of PCWS, as well as the result and analysis of its performance at the Second International Chinese Word Segmentation Bakeoff.

1 Introduction

Och and Ney(2002) present a framework based on direct maximum entropy model to construct the machine translation system. The model treats knowledge sources as feature functions, and allows the system to be extended easily by adding new feature functions. We think the model can be used to provide a unified approach for Chinese word segmentation. PCWS is the system based on this thinking.

2 System Description

PCWS consists of four components: Word generation, Disambiguation, Select the best word sequence and Output the result. They are described below.

2.1 Word Generation

The procedure of word generation involves two steps: (1) generation of the common words which are listed in the Dictionary. (2) generation of the unknown words. The unknown words handled by the system involve numeric expression, time expression, personal name, location name and organization name. PCWS can recognize the abbreviation of person name, and fail to find the abbreviation of location name and organization name.

PCWS constructs an integrated segmentation graph. The node in the graph is the minimal segmentation unit that cannot be split in any stage that follows. The unit consists of Chinese character, punctuation, Arabic numeral string and English character string. Every word that be generated is an edge in the graph.

Making the integrated segmentation graph is to avoid the blind spots in segmentation, but it brings the graph more complex, and make the system's speed slow.

Every word will belong to a class.

Given a word $\mathbf{w}_i$, its class is defined by Figure 1.

$$c_i = \begin{cases} \text{Wi} \quad \text{iff} & \mathbf{wi} \text{ is listed in the segmentation lexicon.} \\ \text{PER iff} & \mathbf{wi} \text{ is a person name} \\ \text{LOC iff} & \mathbf{wi} \text{ is a location name} \\ \text{ORG iff} & \mathbf{wi} \text{ is an organization name} \\ \text{NUM iff} & \mathbf{wi} \text{ is a numeral expression} \\ \text{TIME iff} & \mathbf{wi} \text{ is a time expression} \end{cases}$$

Figure 1: Class Definition of word **wi**

2.2 Disambiguation

In constructing the graph, PCWS detect the ambiguities of the segmentation and classify the ambiguities into two classes: the false ambiguity and the true ambiguity. The former is simply solved by querying a table. The segmentation information around the true ambiguities will be collected and PCWS will give an estimate of each possible segmentation mode of the true ambiguities. These estimates will be used by selecting the best path.

2.3 Select the Best Word Sequence

The procedure of this part consists of two processes: generating the candidate word sequences and finding the best word sequence.

If **S** is a Chinese sentence which is a character sequence, **W** is the all possible word sequences given S, $W^{\#} = \{W_1, W_2, ..., W_N\}$ is a word sequence, $C=\{C_1, C_2, ..., C_N\}$ is a corresponding class sequence of $W^{\#}$. We use Viterbi algorithm to generate the candidate paths. In order to control the search space, all the paths will be ranked by a class mode score. The maximum number of the candidate paths generated by each node in the graph cannot be larger than a Number threshold we give.

The class mode score we used can be written as

$$\mathrm{Score}(w^{\#}) = \mathrm{P}_{\mathrm{generate}}(w^{\#} \mid c)\,\mathrm{P}_{\mathrm{context}}(c)$$

The P(**C**) and P($W^{\#}$ | **C**) is similar to the one defined by Gao et al.(2003).

We use the direct maximum entropy model to find the best word sequence. If W^{+} is the best path we need. **W*** is the candidate set.

Giving the direct maximum entropy model and neglecting its renormalization, we can obtain the following decision rule:

$$w^{+} = \arg\max_{w^{\#} \in w^{*}} \left\{ \sum_{m=1}^{M} \lambda_m \mathrm{h}_{\mathrm{m}} \left(w^{\#}, s \right) \right\}$$

hi (**W**$^{\#}$, **S**) is the feature function of the word sequence. The parameter **λi** is the power of the feature.

In PCWS, we define five feature functions:

1) Context feature

$$\mathrm{h}_1(w^{\#}, s) = \text{-log}\,\mathrm{P}_{\mathrm{context}}(c)$$

2) Candidate person name feature

$$\mathrm{h}_2(w^{\#}, s) = \sum_{\mathrm{i==1}}^{\mathrm{N}} \mathrm{genper}(w_i \mid c_i)$$

Here

$$\mathrm{genper}(w_i|c_i) = \begin{cases} \text{-log}\,\mathrm{P}_{\mathrm{generate}}(w_i|c_i) & \text{iff } c_i == \mathrm{PER} \\ 0 & \text{else} \end{cases}$$

3) Candidate location name feature

$$\mathrm{h}_3(w^{\#}, s) = \sum_{\mathrm{i==1}}^{\mathrm{N}} \mathrm{genloc}(w_i|c_i)$$

Here

$$\mathrm{genloc}(w_i|c_i) = \begin{cases} \text{-log}\mathrm{P}_{\mathrm{generate}}(w_i|c_i) & \text{iff } c_i == \mathrm{LOC} \\ 0 & \text{else} \end{cases}$$

4) Candidate organization name feature

$$\mathrm{h}_4(w^{\#}, s) = \sum_{\mathrm{i==1}}^{\mathrm{N}} \mathrm{genorg}(w_i|c_i)$$

Here

$$\mathrm{genorg}(w_i|c_i) = \begin{cases} \text{-log}\mathrm{P}_{\mathrm{generate}}(w_i|c_i) & \text{iff } c_i == \mathrm{ORG} \\ 0 & \text{else} \end{cases}$$

5) The length of the path

$$\mathrm{h}_5(w^{\#}, s) = \mathrm{Length}(c)$$

We realize the GIS algorithm which can handle any type of real-valued features to train the values of λ_1^5 .

2.4 Output the Result

The component outputs the best word sequence and adjusts the result form based on the standard of test corpora. In PCWS, we only adjust the form of the unknown words the system recognizes.

For example, in "１９９２年１１月", "１９９２年" and "１１月" will be recognized independently as unknown word. Base on the standard of Msr corpora, we combine the continuous words which belong to the class TIME."

2.5 System Overview

The overall architecture of our word segmentation system is presented in figure 2.

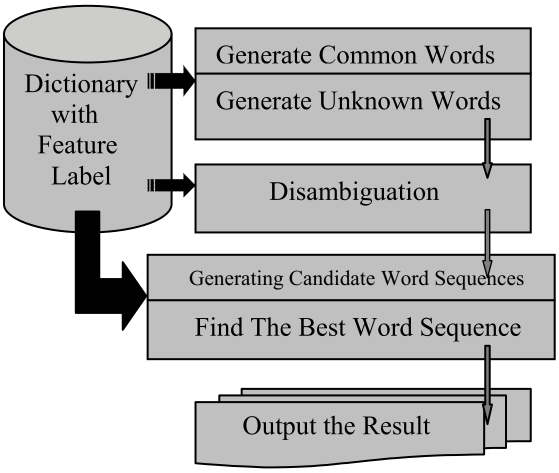
Dictionary with Feature Label
Generate Common Words
Generate Unknown Words
Disambiguation
Generating Candidate Word Sequences
Find The Best Word Sequence
Output the Result

Figure 2: Overall architecture of PCWS

Track	TOTAL TRUE WORD COUNT	TOTAL TEST WORD COUNT	P	R	F	OOV	Roov	Riv
Msr_Open	106873	106624	0.913	0.915	0.914	0.026	0.725	0.918

Table 1: Test Result on Msr-open Track

3 Evaluation

Because of the bakeoff's rule and the limit of time, we only attend the track of Msr_open.
Table 1 is the result of PCWS in this bakeoff.
Form the table, we can know the system is remarkable in OOV recognize.
Due to only small named entity words are included in the PCWS's dictionary, most of named entity words are generated by the system. However, the system's overall performance is not in balance with its good Roov. We notice Riv of the result is low. The main reason causes Riv low is the difference between the segmentation standard of PKU training corpora and the segmentation standard of MSR test corpora. The difference is so distinctly even in the segmentation standard of common words. For instance:

Example 1
[Correct Result] "另外 旅游 、 侨汇 **也是** 经济 收入 的 重要 组成部分 ， 制造业 规模 相对 **较小** 。"
[PCWS Result] "另外 旅游 、 侨汇 **也 是** 经济 收入 的 重要 组成部分 ， 制造业 规模 相对 **较 小** 。"
The result's True Words Recall = 0.875, Test Words Precision = 0.778

Example 2
[Correct Result] "**在此** 之前 的 **回 购** 合同 继续 执行 完毕 。"
[PCWS Result] "**在 此** 之前 的 **回购** 合同 继续 执行 完毕 。"
The result's True Words Recall = 0.700, Test Words Precision = 0.700

Before we get the result, we neglected the problem. We put our attention in named entity recognize, which need the training corpora with the label information, so we use six month PKU corpora to construct our system and not use well the MSR training corpora in the bakeoff.

In order to know the influence of the problem, we test our system in PKU test corpora, Table 2 is the result of PCWS in the test.

Track	P	R	F
Pku_Open	0.973887	0.978405	0.976141

Table 2: Test Result on Pku_open

It's an excellent result and powerful proves our suppose.
We wish to make our system more adaptable to different standards in the near future.

4 Conclusion

We have presented our Chinese word segmentation system PCWS and its result for Msr_open track. We are glad to see its good performance of OOV recognize. In the course of the bakeoff, we find some problems in PCWS. We will try to select more useful feature functions into the existing segmentation model in future work. We are confident the system's performance will have a big progress next time.

5 Acknowledgement

Thank Prof. Hou-Feng Wang for his attention and aid. The author would especially express gratitude to his family, their encouragement and support make the author to hold the line.

References

Franz J. Och and Hermann Ney. 2002. Discriminative training and maximum entropy models for statistical machine translation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), pages 295-302, Philadelphia, PA, July.

Gao, Jianfeng, Mu Li and Chang-Ning Huang. 2003. Improved source-channel model for Chinese word segmentation. In: ACL2003.