Chinese Word Segmentation with Maximum Entropy and N-gram Language Model

Wang Xinhao, Lin Xiaojun, Yu Dianhai, Tian Hao, Wu Xihong
National Laboratory on Machine Perception,
School of Electronics Engineering and Computer Science,
Peking University, China, 100871
{wangxh,linxj,yudh,tianhao,wxh}@cis.pku.edu.cn

Abstract

This paper presents the Chinese word segmentation systems developed by Speech and Hearing Research Group of National Laboratory on Machine Perception (NLMP) at Peking University, which were evaluated in the third International Chinese Word Segmentation Bakeoff held by SIGHAN. The Chinese character-based maximum entropy model, which switches the word segmentation task to a classification task, is adopted in system developing. To integrate more linguistics information, an n-gram language model as well as several post processing strategies are also employed. Both the closed and open tracks regarding to all four corpora MSRA, UPUC, CITYU, CKIP are involved in our systems' evaluation, and good performance are achieved. Especially, in the closed track on MSRA, our system ranks 1st.

1 Introduction

Chinese word segmentation is one of the core techniques in Chinese language processing and attracts lots of research interests in recent years. Several promising methods are proposed by previous researchers, in which Maximum Entropy (ME) model has turned out to be a successful way for this task (Hwee Tou Ng et al., 2004; Jin Kiat Low et al., 2005). By employing Maximum Entropy (ME) model, the Chinese word segmentation task is regarded as a classification problem, where each character will be classified to one of the four classes, i.e., the *beginning*, *middle*, *end* of a multi-character word and a single-character word.

However, in a high degree, ME model pays its emphasis on Chinese characters while debases the consideration on the relationship of the context words. Motivated by this view, several strategies used for reflecting the context words' relationship and integrating more linguistics information, are employed in our systems.

As known, an n-gram language model could express the relationship of the context words well, it therefore as a desirable choice is imported in our system to modify the scoring of the ME model. An analysis on our preliminary experiments shows the combination ambiguity is another issue that should be specially tackled, and a division and combination strategy is then adopted in our system. To handle the numeral words, we also introduce a number conjunction strategy. In addition, to deal with the long organization names problem in MSRA corpus, a post processing strategy for organization name is presented.

The remainder of this paper is organized as follows. Section 2 describes our system in detail. Section 3 presents the experiments and results. And in last section, we draw our conclusions.

2 System Description

With the ME model, n-gram language model, and several post processing strategies, our systems are established. And detailed description on these components are given in following subsections.

2.1 Maximum Entropy Model

The ME model used in our system is based on the previous works (Jin Kiat Low et al., 2005; Hwee Tou Ng et al., 2004). As mentioned above, the ME model based word segmentation is a 4-classes learning process. Here, we remarked four classes, i.e. the beginning, middle, end of a multi-character

Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing, pages 138–141,
Sydney, July 2006.

word and a single-character word, as b, m, e and s respectively.

In ME model, the following features (Jin Kiat Low et al., 2005) are selected:

a) c_n $(n = -2, -1, 0, 1, 2)$

b) $c_n c_{n+1}$ $(n = -2, -1, 0, 1)$

c) $c_{-1}c_{+1}$

where c_n indicates the character in the left or right position n relative to the current character c_0.

For the open track especially, three extended features are extracted with the help of an external dictionary as follows:

d) Pu (c_0)

e) L and t_0

f) $c_n t_0$ $(n = -1, 0, 1)$

where Pu(c_0) denotes whether the current character is a punctuation, L is the length of word *W* that conjoined from the character and its context which matching a word in the external dictionary as long as possible. t_0 is the boundary tag of the character in *W*.

With the features, a ME model is trained which could output four scores for each character with regard to four classes. Based on scores of all characters, a completely segmented semiangle matrix can be constructed. Each element w_{ji} in this matrix represents a word that starts at the ith character and ends at jth character, and its value $\texttt{ME}(j, i)$, the score for these $(j - i + 1)$ characters to form a word, is calculated as follow:

$$\begin{aligned} \texttt{ME}[j,i] &= -\log p(w = c_i...c_j) \\ &= -\log[p(b_{c_i})p(m_{c_{i+1}})... \\ &\quad p(m_{c_{j-1}})p(e_{c_j})] \end{aligned} \quad (1)$$

As a consequence, the optimal segmentation results corresponding to the best path with the lowest overall score could be reached via a dynamic programming algorithm. For example:

那一年我十九岁(I was 19 years old that year)

Table 1 shows its corresponding matrix. In this example, the ultimate segmented result is:

那 一年 我 十九岁

2.2 Language Model

N-gram language model, a widely used method in natural language processing, can represent the context relation of words. In our systems, a bigram model is integrated with ME model in the phase of calculating the path score. In detail, the score of a path will be modified by adding the bigram of words with a weight λ at the word boundaries. The approach used for modifying path score is based on the following formula.

$$\begin{aligned} V[j,i] &= \texttt{ME}[j,i] \\ &\quad + \min_{k=1}^{i-1}\{[(V[i-1,k] \\ &\quad + \lambda \texttt{Bigram}(w_{k,i-1}, w_{i,j})\} \end{aligned} \quad (2)$$

where V[j,i] is the score of local best path which ends at the jth character and the last word on the path is $w_{i,j} = c_i...c_j$, the parameter λ is optimized by the test set used in the 2nd International Chinese Word Segmentation Bakeoff. When scoring the path, if one of the words $w_{k,i-1}$ and $w_{i,j}$ is out of the vocabulary, their bigram will backoff to the unigram. And the unigram of the OOV word will be calculated as:

$$\texttt{Unigram}(\texttt{OOV Word}) = p^l \quad (3)$$

where *p* is the minimal unigram value of words in vocabulary; *l* is the length of the word acting as a punishment factor to avoid overemphasizing the long OOV words.

2.3 Post Processing Strategies

The analysis on preliminary experiments, where the ME model and n-gram language model are involved, lead to several post processing strategies in developing our final systems.

2.3.1 Division and Combination Strategy

To handle the combination ambiguity issue, we introduce a division and combination strategy which take in use of unigram and bigram. For each two words A and B, if their bigrams does not exist while there exists the unigram of word AB, then they can be conjoined as one word. For example, "十月(August)" and "革命(revolution)" are two segmented words, and in training set the bigram of "十 月" and "革 命" is absent, while the word "十月革命(the August Revolution)" appeares, then the character string "十月革命" is conjoined as one word. On the other hand, for a word C which can be divided as AB, if its unigram does not exit in training set, while the bigram of its subwords A and B exits, then it will be resegmented. For example, Taking the word "经济体制改革(economic system reform)" for instance, if its corresponding unigram is absent in training set, while the bigram of two subwords "经济 体

		那	一	年	我	十	九	岁
		1	2	3	4	5	6	7
那	1	**6.3180e-07**						
一	2	33.159	7.5801					
年	3	26.401	**0.0056708**	5.2704				
我	4	71.617	45.221	49.934	**3.1001e-07**			
十	5	83.129	56.734	61.446	33.869	7.0559		
九	6	90.021	63.625	68.337	40.760	12.525	12.534	
岁	7	77.497	51.101	55.813	28.236	**0.0012012**	10.077	10.055

Table 1: A completely segmented matrix

制(economic system)" and "改革(reform)" exists, as a consequence, it will be segmented into two words "经济体制" and "改革".

2.3.2 Numeral Word Processing Strategy

The ME model always segment a numeral word into several words. For instance, the word "4.34元(RMB Yuan 4.34)", may be segmented into two words "4." and "34元". To tackle this problem, a numeral word processing strategy is used. Under this strategy, those words that contain Arabic numerals are manually marked in the training set firstly, then a list of high frequency characters which always appear alone between the numbers in the training set can be extracted, based on which numeral word issue can be tackled as follows. When segmenting one sentence, if two conjoint words are numeral words, and the last character of the former word is in the list, then they are combined as one word.

2.3.3 Long Organization Name Processing Strategy

Since an organization name is usually an OOV, it always will be segmented as several words, especially for a long one, while in MSRA corpus, it is required to be recognized as one word. In our systems, a corresponding strategy is presented to deal with this problem. Firstly a list of organization names is manually selected from the training set and stored in the prefix-tree based on characters. Then a list of prefixes is extracted by scanning the prefix-tree, that is, for each node, if the frequencies of its child nodes are all lower than the predefined threshold *k* and half of the frequency of the current node, the string of the current node will be extracted as a prefix; otherwise, if there exists a child node whose frequency is higher than the threshold *k*, scan the corresponding subtree. In the same way, the suffixes can also be extracted. The only difference is that the order of characters is inverse in the lexical tree.

During recognizing phase, to a successive words string that may include 2-5 words, will be combined as one word, if all of the following conditions are satisfied.

a) Does not include numbers, full stop or comma.

b) Includes some OOV words.

c) Has a tail substring matching some suffix.

d) Appears more than twice in the test data.

e) Has a higher frequency than any of its substring which is an OOV word or combined by multiple words.

f) Satisfy the condition that for any two successive words w_1 w_2 in the strings, freq(w_1w_2)/freq(w_1)$\geq$0.1, unless w_1 contains some prefix in its right.

3 Experiments and Results

We have participated in both the closed and open tracks of all the four corpora. For MSRA corpus and other three corpora, we build System I and System II respectively. Both systems are based on the ME model and the Maximum Entropy Toolkit [1], provided by Zhang Le, is adopted.

Four systems are derived from System I with regard to whether or not the n-gram language model and three post processing strategies are used on the closed track of MSRA corpus. Table 2 shows the results of four derived systems.

System	R	P	F	R_{OOV}	R_{IV}
IA	95.0	95.7	95.3	66.0	96.0
IB	96.0	95.6	95.8	60.3	97.3
IC	96.4	96.0	96.2	60.3	97.7
ID	96.4	96.1	96.3	61.2	97.6

Table 2: The effect of ME model, n-gram language model and three post processing strategies on the closed track of MSRA corpus.

System **IA** only adopts the ME model. System **IB** integrates the ME model and the bigram language model. System **IC** integrates the division and combination strategy and the numeral words

[1] http://homepages.inf.ed.ac.uk/s0450736/maxent_toolkit.html

processing strategy. System **ID** adds the long organization name processing strategy.

For the open track of MSRA, an external dictionary is utilized to extract the *e* and *f* features. The external dictionary is built from six sources, including the Chinese Concept Dictionary from Institute of Computational Linguistics, Peking University(72,716 words), the LDC dictionary(43,120 words), the Noun Cyclopedia(111,633), the word segmentation dictionary from Institute of Computing Technology, Chinese Academy of Sciences(84,763 words), the dictionary from Institute of Acoustics, and the dictionary from Institute of Computational Linguistics, Peking University(68,200 words) and a dictionary collected by ourselves(63,470 words).

The union of the six dictionaries forms a *big dictionary*, and those words appearing in five or six dictionaries are extracted to form a *core dictionary*. If a word belongs to one of the following dictionaries or word sets, it is added into the external dictionary.

a) The core dictionary.

b) The intersection of the big dictionary and the training data.

c) The words appearing in the training data twice or more times.

Those words in the external dictionaries will be eliminated, if in most cases they are divided in the training data. Table 3 shows the effect of ME model, n-gram language model, three post processing strategies on the open track of MSRA. Here System IO only adopts the basic features, while the external dictionary based features are used in four derived systems related to open track: IA, IB, IC, ID.

System	R	P	F	R_{OOV}	R_{IV}
IO	96.0	96.5	96.3	71.1	96.9
IA	97.5	96.9	97.2	65.9	98.6
IB	97.6	96.8	97.2	64.8	98.7
IC	97.7	97.0	97.4	66.8	98.8
ID	97.7	97.1	97.4	67.5	98.8

Table 3: The effect of ME model, n-gram language model, three post processing strategies on the open track of MSRA.

System II only adopts ME model, the division and combination strategy and the numeral word processing strategy. In the open track of the corpora CKIP and CITYU, the training set and test set from the 2nd Chinese Word Segmentation Backoff are used for training. For the corpora UPUC and CITYU, the external dictionaries are used, which is constructed in the same way as that in the open track of MSRA Corpus. Table 4 shows the official results of system II on UPUC, CKIP and CITYU.

Corpus	R	P	F	R_{OOV}	R_{IV}
UPUC-C	93.6	92.3	93.0	68.3	96.1
UPUC-O	94.0	90.7	92.3	56.1	97.6
CKIP-C	95.8	94.8	95.3	64.6	97.2
CKIP-O	95.8	94.8	95.3	64.7	97.2
CITYU-C	96.9	97.0	97.0	77.3	97.8
CITYU-O	97.9	97.6	97.7	81.3	98.5

Table 4: Official results of our systems on UPUC CKIP and CITYU

On the UPUC corpus, an interesting observation is that the performance of the open track is worse than the closed track. The investigation and analysis lead to a possible explanation. That is, the segmentation standard of the dictionaries, which are used to construct the external dictionary, is different from that of the UPUC corpus.

4 Conclusion

In this paper, a detailed description on several Chinese word segmentation systems are presented, where ME model, n-gram language model as well as three post processing strategies are involved. In the closed track of MSRA, the integration of bigram language model greatly improves the recall ratio of the words in vocabulary, although it will impairs the performance of system in recognizing the words out of vocabulary. In addition, three strategies are introduced to deal with combination ambiguity, numeral word, long organization name issues. And the evaluation results reveal the validity and effectivity of our approaches.

References

Jin Kiat Low, Hwee Tou Ng and Wenyuan Guo. A maximum Entropy Approach to Chinese Word Segmentation. 2005. *Preceedings of the Fourth SIGHAN Workshop on Chinese Language Processing*, pp. 161-164.

Hwee Tou Ng and Jin Kiat Low. Chinese part-of-speech tagging: One-at-a-time or all-at-once? word-based or character-based? 2004. *Preceedings of the 2004 Conference on Empirical Methods in Natural Language Processing(EMNLP)*, pp. 277-284.

Zhang Huaping and Liu Qun. Model of Chinese Words Rough Segmentation Based on N-Shortest-Paths Method. 2002. *Journal of Chinese Information Processing*, 28(1):pp. 1-7.