A Pragmatic Chinese Word Segmentation System

Wei Jiang, Yi Guan, Xiao-Long Wang
School of Computer Science and Technology, Harbin Institute of Technology,
Heilongjiang Province, 150001, P.R.China
jiangwei@insun.hit.edu.cn

Abstract

This paper presents our work for participation in the Third International Chinese Word Segmentation Bakeoff. We apply several processing approaches according to the corresponding sub-tasks, which are exhibited in real natural language. In our system, Trigram model with smoothing algorithm is the core module in word segmentation, and Maximum Entropy model is the basic model in Named Entity Recognition task. The experiment indicates that this system achieves F-measure 96.8% in MSRA open test in the third SIGHAN-2006 bakeoff.

1 Introduction

Word is a logical semantic and syntactic unit in natural language. Unlike English, there is no delimiter to mark word boundaries in Chinese language, so in most Chinese NLP tasks, word segmentation is a foundation task, which transforms Chinese character string into word sequence. It is prerequisite to POS tagger, parser or further applications, such as Information Extraction, Question Answer system.

Our system participated in the Third International Chinese Word Segmentation Bakeoff, which held in 2006. Compared with our system in the last bakeoff (Jiang 2005A), the system in the third bakeoff is adjusted intending to have a better pragmatic performance. This paper mainly focuses on describing two sub-tasks: (1) The basic Word Segmentation; (2) Named entities recognition. We apply different approaches to solve above two tasks, and all the modules are integrated into a pragmatic system (**ELUS)**.

2 System Description

All the words in our system are categorized into five types: Lexicon words (LW), Factoid words (FT), Morphologically derived words (MDW), Named entities (NE), and New words (NW). Figure 1 demonstrates our system structure.

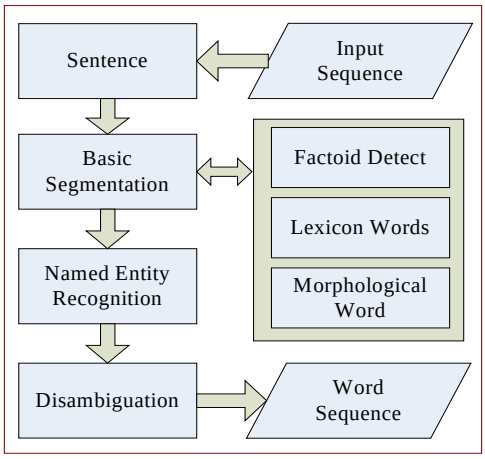

Figure 1 ELUS Segmenter and NER

The input character sequence is converted into one or several sentences, which is the basic dealing unit. The "Basic Segmentation" is used to identify the LW, FT, MDW words, and "Named Entity Recognition" is used to detect NW words. We don't adopt the New Word detection algorithm in our system in this bakeoff. The "Disambiguation" module performs to classify complicated ambiguous words, and all the above results are connected into the final result, which is denoted by XML format.

2.1 Trigram and Smoothing Algorithm

We apply the trigram model to the word segmentation task (Jiang 2005A), and make use of Absolute Smoothing algorithm to overcome the sparse data problem.

Trigram model is used to convert the sentence into a word sequence. Let $\mathbf{w} = w_1 w_2 \ldots w_n$ be a word sequence, then the most likely word sequence w* in trigram is:

$$\mathbf{w}^* = \arg\max_{w_1 w_2 \cdots w_n} \prod_{i=1}^{n} P(w_i \mid w_{i-2} w_{i-1}) \qquad (1)$$

where let $P(w_0|w_{-2}\ w_{-1})$ be $P(w_0)$ and let $P(w_1|w_{-1}\ w_0)$ be $P(w_1|w_0)$, and w_i represents LW or a type of FT or MDW. In order to search the best segmentation way, all the word candidates are filled in the word lattice (Zhao 2005). And the Viterbi

Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing, pages 189–192,
Sydney, July 2006.

algorithm is used to search the best word segmentation path.

FT and MDW need to be detected when constructing word lattice (detailed in section 2.2). The data structure of lexicon can affect the efficiency of word segmentation, so we represent lexicon words as a set of TRIEs, which is a tree-like structure. Words starting with the same character are represented as a TRIE, where the root represents the first Chinese character, and the children of the root represent the second characters, and so on (Gao 2004).

When searching a word lattice, there is the zero-probability phenomenon, due to the sparse data problem. For instance, if there is no cooccurence pair "我们/吃/香蕉"(we eat bananas) in the training corpus, then P(香蕉|我们，吃) = 0. According to formula (1), the probability of the whole candidate path, which includes "我们/吃/香蕉" is zero, as a result of the local zero probability. In order to overcome the sparse data problem, our system has applied Absolute Discounting Smoothing algorithm (Chen, 1999).

$$N_{1+}(w_{i-n+1}^{i-1}\bullet) = |\{w_i : c(w_{i-n+1}^{i-1}w_i) > 0\}| \quad (2)$$

The notation N_{1+} is meant to evoke the number of words that have one or more counts, and the $\bullet$ is meant to evoke a free variable that is summed over. The function $c()$ represents the count of one word or the cooccurence count of multi-words. In this case, the smoothing probability

$$p(w_i \mid w_{i-n+1}^{i-1}) = \frac{\max\{c(w_{i-n+1}^{i}) - D, 0\}}{\sum_{w_i} c(w_{i-n+1}^{i})} + (1-\lambda)\, p(w_i \mid w_{i-n+2}^{i-1}) \quad (3)$$

$$\text{where, } 1-\lambda = \left(\frac{D}{\sum_{w_i} c(w_{i-n+1}^{i})} N_{1+}(w_{i-n+1}^{i-1}\bullet) \right) \quad (4)$$

Because we use trigram model, so the maximum n may be 3. A fixed discount D ($0 \le D \le 1$) can be set through the deleted estimation on the training data. They arrive at the estimate

$$D = \frac{n_1}{n_1 + 2n_2} \quad (5)$$

where n_1 and n_2 are the total number of n-grams with exactly one and two counts, respectively.

After the basic segmentation, some complicated ambiguous segmentation can be further disambiguated. In trigram model, only the previous two words are considered as context features, while in disambiguation processing, we can use the Maximum Entropy model fused more features (Jiang 2005B) or rule based method.

2.2 Factoid and Morphological words

All the Factoid words can be represented as regular expressions. So the detection of factoid words can be achieved by Finite State Automaton(FSA). In our system, the following categories of factoid words can be detected, as shown in table 1.

Table 1 Factoid word categories

FT type	Factoid word	Example
Number	Integer, real, percent etc.	2910, 46.12%, 二十九, 三千七百二十
Date	Date	2005 年 5 月 12 日
Time	Time	8:00, 十点二十分
English	English word,	How, are, you
www	Website, IP address	http://www.hit.edu.cn 192.168.140.133
email	Email	elus@google.com
phone	Phone, fax	0451-86413322

Deterministic FSA (DFA) is efficient because a unique "next state" is determined, when given an input symbol and the current state. While it is common for a linguist to write rule, which can be represented directly as a non-deterministic FSA (NFA), i.e. which allows several "next states" to follow a given input and state.

Since every NFA has an equivalent DFA, we build a FT rule compiler to convert all the FT generative rules into a DFA. e.g.

- "< digit > -> [0..9];
- < year > ::= < digit >{< digit >+}年";
- < integer > ::= {< digit >+};

where "->" is a temporary generative rule, and "::=" is a real generative rule.

As for the morphological words, we erase the dealing module, because the word segmentation definition of our system adopts the PKU standard.

3 Named Entity Recognition

We adopt Maximum Entropy model to perform the Named Entity Recognition. The extensive evaluation on NER systems in recent years (such as CoNLL-2002 and CoNLL-2003) indicates the best statistical systems are typically achieved by using a linear (or log-linear) classification algorithm, such as Maximum Entropy model, together with a vast amount of carefully designed linguistic features. And this seems still true at present in terms of statistics based methods.

Maximum Entropy model (ME) is defined over $H \times T$ in segmentation disambiguation, where H is the set of possible contexts around target word that will be tagged, and T is the set of allowable tags, such as B-PER, I-PER, B-LOC, I-LOC etc. in our NER task. Then the model's conditional probability is defined as

$$p(t \mid h) = \frac{p(h,t)}{\sum_{t' \in T} p(h,t')} \quad (6)$$

where $p(h,t) = \pi\mu \prod_{j=1}^{k} \alpha_j^{f_j(h,t)}$ (7)

where h is the current context and t is one of the possible tags.

The several typical kinds of features can be used in the NER system. They usually include the context feature, the entity feature, and the total resource or some additional resources.

Table 2 shows the context feature templates.

Table2 NER feature template[1]

Type	Feature Template
One order feature	w_{i-2}, w_{i-1}, w_i, w_{i+1}, w_{i+2}
Two order feature	$w_{i-1:i}$, $w_{i:i+1}$
NER tag feature	t_{i-1}

While, we only point out the local feature template, some other feature templates, such as long distance dependency templates, are also helpful to NER performance. These trigger features can be collected by Average Mutual Information or Information Gain algorithm etc.

Besides context features, entity features is another important factor, such as the suffix of Location or Organization. The following 8 kinds of dictionaries are usually useful (Zhao 2006):

Table 3 NER resource dictionary[2]

List Type	Lexicon	Example
Word list	Place lexicon	北京, 纽约, 马家沟
	Chinese surname	张, 王, 赵, 欧阳
String list	Prefix of PER	老, 阿, 小
	Suffix of PLA	山, 湖, 寺, 台, 海
	Suffix of ORG	会, 联盟, 组织, 局
Character list	Character for CPER	军, 刚, 莲, 茵, 倩
	Character for FPER	科, 曼, 斯, 娃, 贝
	Rare character	滢, 胨, 蓠

In addition, some external resources may improve the NER performance too, e.g. we collect a lot of entities for Chinese Daily Newspaper in 2000, and total some entity features.

However, our system is based on Peking University (PKU) word segmentation definition and PKU NER definition, so we only used the basic features in table 2 in this bakeoff. Another effect is the corpus: our system is training by the Chinese Peoples' Daily Newspaper corpora in 1998, which conforms to PKU NER definition. In the section 4, we will give our system performance with the basic features in Chinese Peoples' Daily Newspaper corpora.

[1] w_i – current word, w_{i-1} – previous word, t_i – current tag.

[2] Partial translation: 北京 BeiJing,纽约 New york;张 Zhang, 王 Wang; 老 Old;山 mountain,湖 lake;局 bureau.

4 Performance and analysis

4.1 The Evaluation in Word Segmentation

The performance of our system in the third bakeoff is presented in table 4 in terms of recall(R), precision(P) and F score in percentages. The score software is standard and open by SIGHAN.

Table 4 MSRA test in SIGHAN2006 (%)

MSRA	R	P	F	OOV	R_{oov}	R_{iv}
Close	96.3	91.8	94.0	3.4	17.5	99.1
Open	97.7	96.0	96.8	3.4	62.4	98.9

Our system has good performance in terms of R_{iv} measure. The R_{iv} measure in close test and in open test are 99.1% and 98.9% respectively. This good performance owes to class-based trigram with the absolute smoothing and word disambiguation algorithm.

In our system, it is the following reasons that the open test has a better performance than the close test:

(1) Named Entity Recognition module is added into the open test system. And Named Entities, including PER, LOC, ORG, occupy the most of the out-of-vocabulary words.

(2) The system of close test can only use the dictionary that is collected from the given training corpus, while the system of open test can use a better dictionary, which includes the words that exist in MSRA training corpus in SIGHAN2005. And we know, the dictionary is the one of important factors that affects the performance, because the LW candidates in the word lattice are generated from the dictionary.

As for the dictionary, we compare the two collections in SIGHAN2005 and SIGHAN2006, and evaluating in SIGHAN2005 MSRA close test. There are less training sentence in SIGHAN2006, as a result, there is at least 1.2% performance decrease. So this result indicates that the dictionary can bring an important impact in our system.

Table 5 gives our system performance in the second bakeoff. We'll make brief comparison.

Table 5 MSRA test in SIGHAN 2005 (%)

MSRA	R	P	F	OOV	R_{oov}	R_{iv}
Close	97.3	94.5	95.9	2.6	32.3	99.1
Open	98.0	96.5	97.2	2.6	59.0	99.0

Comparing table 4 with table 5, we find that the OOV is 3.4 in third bakeoff, which is higher than the value in the last bakeoff. Obviously, it is one of reasons that affect our performance.

In addition, based on pragmatic consideration, our system has been made some simplifier, for instance, we erase the new word detection algorithm and the is no morphological word detection.

4.2 Named Entity Recognition

In MSRA NER open test, our NER system is training in prior six-month corpora of Chinese Peoples' Daily Newspaper in 1998, which were annotated by Peking University. Table 6 shows the NER performance in the MSRA open test.

Table 6 The NER performance in MSRA Open test

MSRA NER	Precision	Recall	F Score
PER	93.68%	86.37%	89.87
LOC	85.50%	59.67%	70.29
ORG	75.87%	47.48%	58.41
Overall	86.97%	65.56%	74.76

As a result of insufficiency in preparing bakeoff, our system is only trained in Chinese Peoples' Daily Newspaper, in which the NER is defined according to PKU standard. However, the NER definition of MSRA is different from that of PKU, e.g, "中华/LOC 民族", "马/PER 列/PER 主义" in MSRA, are not entities in PKU. So the training corpus becomes a main handicap to decrease the performance of our system, and it also explains that there is much difference between the recall rate and the precision in table 6.

Table 7 gives the evaluation of our NER system in Chinese Peoples' Daily Newspaper, training in prior five-month corpora and testing in the sixth month corpus. We also use the feature templates in table 2, in order to make comparison with table 6.

Table 7 The NER test in Chinese Peoples' Daily

MSRA NER	Precision	Recall	F Score
CPN	93.56	90.96	92.24
FPN	90.42	86.47	88.40
LOC	91.94	90.52	91.22
ORG	88.38	84.52	86.40
Overall	91.35	88.85	90.08

This experiment indicates that our system can have a good performance, if the test corpus and the training corpora conform to the condition of independent identically distributed attribution.

4.3 Analysis and Discussion

Some points need to be further considered:

(1) The dictionary can bring a big impact to the performance, as the LW candidates come from the dictionary. However a big dictionary can be easily acquired in the real application.

(2) Due to our technical and insufficiently preparing problem, we use the PKU NER definition, however they seem not unified with the MSRA definition.

(3) Our NER system is a word-based model, and we have find out that the word segmentation with two different dictionaries can bring a big impact to the NER performance.

(4) We erase the new word recognition algorithm in our system. While, we should explore the real annotated corpora, and add new word detection algorithm, if it has positive effect. e.g. "荷花 奖"(lotus prize) can be recognized as one word by the conditional random fields model.

5 Conclusion

We have briefly described our word segmentation system and NER system. We use word-based features in the whole processing. Our system has a good performance in terms of R_{iv} measure, so this means that the trigram model with the smoothing algorithm can deal with the basic segmentation task well. However, the result in the bakeoff indicates that detecting out-of-vocabulary word seems to be a harder task than dealing with the segmentation-ambiguity task.

The work in the future will concentrate on two sides: improving the NER performance and adding New Word Detection Algorithm.

References

HuaPing Zhang, Qun Liu etc. 2003. Chinese Lexical Analysis Using Hierarchical Hidden Markov Model, Second SIGHAN workshop affiliated with 4th ACL, Sapporo Japan, pp.63-70.

Jianfeng Gao, Mu Li et al. 2004. Chinese Word Segmentation: A Pragmatic Approach. MSR-TR-2004-123, November 2004.

Peng Fuchun, Fangfang Feng and Andrew McCallum. Chinese segmentation and new word detection using conditional random fields. In:COLING 2004.

Stanley F.Chen and J. Goodman. 1999. An empirical study of smoothing techniques for language modeling. Computer Speech and Language. 13:369-394.

Wei Jiang, Jian Zhao et al. 2005A.Chinese Word Segmentation based on Mixing Model. 4th SIGHAN Workshop. pp. 180-182.

Wei Jiang, Xiao-Long Wang, Yi Guan et al. 2005B. applying rough sets in word segmentation disambiguation based on maximum entropy model. Journal of Harbin Institute of Technology (New Series). 13(1): 94-98.

Zhao Jian. 2006. Research on Conditional Probabilistic Model and Its Application in Chinese Named Entity Recognition. Ph.D. Thesis. Harbin Institute of Technology, China.

Zhao Yan. 2005. Research on Chinese Morpheme Analysis Based on Statistic Language Model. Ph.D. Thesis. Harbin Institute of Technology, China.