Word Segmentation and Named Entity Recognition for SIGHAN Bakeoff3

Zhang Suxiang
CISTR,
Beijing University of Posts and Telecommunications
zsuxiang@163.com

Qin Ying
CISTR,
Beijing University of Posts and Telecommunications
qinyingmail@163.com

Wen Juan
CISTR,
Beijing University of Posts and Telecommunications
mystery999@163.com

Wang Xiaojie
CISTR,
Beijing University of Posts and Telecommunications
xjwang@bupt.edu.cn

Abstract

We have participated in three open tracks of Chinese word segmentation and named entity recognition tasks of SIGHAN Bakeoff3. We take a probabilistic feature based Maximum Entropy (ME) model as our basic frame to combine multiple sources of knowledge. Our named entity recognizer achieved the highest F measure for MSRA, and word segmenter achieved the medium F measure for MSRA. We find effective combining of the external multi-knowledge is crucial to improve performance of word segmentation and named entity recognition.

1 Introduction

Word Segmentation (WS) and Named Entity Recognition (NER) are two basic tasks for Chinese Processing. The main difficulty is ambiguities widely exist in these two tasks. Our system is thus pay special attentions on various ambiguities resolution. After preprocessing we take Maximum Entropy (ME) model as the unified frame for WS and NER. ME is a effective model which often used to combine multiple sources of knowledge into various features. For finer-grain utilization of features, we use probabilistic features instead of binary features normally used. By exploring some often used features and some new features, our system performs well in this SIGHAN contest.

In the rest sections of this paper, we give a brief introduction to our system sequentially. Section 2 describes the preprocessing in the system, including rough segmentation and factoid identification. Section 3 is on ambiguity resolution of WS. NER is introduced in Section 4. We give some experimental results in Section 5. Finally we draw some conclusions.

2 Preprocessing

The first step in preprocessing is to do a rough segmentation. By using both Forward Maximum Matching (FMM) and Backward Maximum Matching (BMM) approaches, we get an initial segmentation simultaneously detecting some of segmentation ambiguities in text. We use two different wordlists in this step. One is a basic wordlists with about 60 thousands words. We think this wordlist is relatively steady in Chinese. Another includes some words from special training corpus.

We then cope with factoid recognition by using automata. Four automata are built to identify time, date, number and other (like telephone number and model of product) respectively. For covering some exceptional structures, we use some templates to post-process some outputs from automata.

Overlapping and combination ambiguities detected in preprocessing will be treated in next round of our system. It is the topic of next section.

3 Disambiguation

3.1 Overlapping ambiguity

We only detect overlapping ambiguity with length of chain no more than 3 because these kinds of overlapping account for over 98% of all occurrences according to (Yan, 2000). The class-based bigram model trained on tagged corpus of People's Daily 2000 (about 12 million Chinese characters) is applied to resolve the ambiguities. In class-based bigram, all named entities, all punctuation and factoids is one class separately and each word is one class. For MSRA test we

Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing, pages 158–161,
Sydney, July 2006.

evaluate the performance of our overlapping disambiguation with precision of 84.1%.

3.2 Combination ambiguity

We use some templates to describe the POS properties of combination ambiguity and their segmentation words. In our system there are 155 most frequent combination words. Due to the fact that instances of combination ambiguity is deficient in given training corpus, to enlarge training examples we convert the People Daily 2000 to meet the standard of different guidelines then extract examples for training besides the given training corpora. For example, 结果 is a combination ambiguity according to the guideline of MSRA whereas it is always one unit in People Daily 2000. Noticing that when 结果 takes the sense of result, it is always tagged as a noun and a verb when it takes the meaning of fructification, we can easily enlarge the training examples of 结果.

We then use ME model to combination ambiguity resolution. There are six features used in the model as below.

(1) Contextual words;

(2) Contextual characters;

(3) Bigram collocations;

(4) If the transfer probability of adjacent words to the target word exists;

(5) If keywords indicate segmentation exists;

(6) The most frequent segmentation from prior distribution

4 Named entity recognition

4.1 Personal name recognition

We propose a probabilistic feature based maximum entropy approach to NER. Where, probabilistic feature functions are used instead of binary feature functions, it is one of the several differences between this model and the most of the previous ME based model. We also explore several new features in our model, which includes confidence functions, position of features etc. Like those in some previous works, we use sub-models to model Chinese Person Names, Foreign Names respectively, but we bring some new techniques in these sub-models.

In standard ME, feature function is a binary function, for example, if we use CPN denotes the Chinese person Name, SN denotes Surname, a typical feature is:

$$f_i(x,y)=\begin{cases}1 & y\in CPN \quad and \quad x\in SN \\ 0 & otherwise\end{cases} \qquad (1)$$

But in Chinese, firstly, most of words used as surname are also used as normal words. The probabilities are different for them to be used as surname. Furthermore, a surname is not always followed by a given name, both cases are not binary. To model these phenomena, we give probability values to features, instead of binary values.

For example, a feature function can be set value as follows:

$$f(x,y)=\begin{cases}0.985 & \text{if } y\in\text{CPN and } x\in\text{郭} \\ 0 & otherwise\end{cases} \qquad (2)$$

Or

$$f(x,y)=\begin{cases}0.01805 & \text{if } y\in\text{CPN } and \text{ x}\in\text{于} \\ 0 & otherwise\end{cases} \qquad (3)$$

Chinese characters used for translating foreign personal name are different from those in Chinese personal name. We built the foreign name model by collecting suffixes, prefixes, frequently-used characters, estimate their probabilities used in foreign personal name. These probabilities also used in model as probability features.

We also design a confidence function for a character sequence $W=C_1C_2...C_n$ to help model to estimate the probability of W as a person name. C_i may be a character or a word. Let f_{1F} is probability of the C_1, f_{iM} is the probability of the C_i , f_{nE} is the probability of the C_n. So the confidence function is

$$K(w,PERSON)=f_{1F}+\sum_{2<=i<=n-1}f_{iM}+f_{nE} \qquad (4)$$

This function is included in ME frame as a feature.

Candidate person name collection is the first step of NER. Since the ambiguity of Chinese word segmentation always exists. We propose some patterns for model different kind of segmentation ambiguity. Some labels are used to express specific roles of Chinese characters in person names.

We have seven patterns as follows; first two patterns are non-ambiguity, while the others model some possible ambiguity in Chinese person name brought by word segmenter.

(1) BCD: the Chinese personal name is composed of three Hanzi ((Chinese character).

B: Surname of a Chinese personal name.

C: Head character of 2-Hanzi given names.

D: Tail character of 2-Hanzi of given names.

(2) BD: the Chinese personal name is composed of two Hanzi (Chinese character).

(3) BCH:

H: the last given name and its next context are composed of a word.

(4) UCD:

U: the surname and its previous context are composed of a word.

(5) BE:

E: the first given name and the last given name are composed of a word.

(6) UD:

U: the surname and the first given name are composed of a word.

(7) CD：The Chinese personal name is only composed of two given names.

Based on the People's Daily corpus and maximum entropy, we achieve models of Chinese personal name and transliterated personal name respectively.

Here, How can we know whether a person name is composed of two or three Hanzi, we used another technology to limit boundary. We think out the co-appearing about the last given name and its next context, now, we have made a statistics about personal name and its next context to decide the length of the Chinese personal name. For example:

"李超为宁波拿下了一分",

In this sentence, we collect a candidate Chinese person name "李超为", but the last given name "为" is a specific character, it has different meaning, now, we make a decision whether "为" is belong to personal name or not.

$$number(NR \quad 宁波) < number(NR \quad 为) \quad (3)$$

So, "为" is not included in the personal name, "李超" is a correct choice.

Another problem we have met is to recognize transliterated personal name, because many transliterated personal characters has included the Chinese surname, however, the condition that we can recognize the Chinese personal name is Chinese surname, therefore, a section of the transliterated personal name will often be recognized a Chinese personal name.

In our system, we design a dynamitic priority method to check ambiguous character, when we examine a ambiguous character like "谢" or "马", we will search different characters which maybe belong to Chinese personal name or transliterated personal name with forward and backward direction. According to the collection result, we will decide to use Chinese personal model or transliterated personal model to recognize personal name.

For example:

"印/方/重工业/和/国营/企业/部/部长/马/诺/哈/尔/·/乔/希/、/随/访/的/部分/议员/以及/印度/驻/华/大使/南/威/哲/等/参加/了会见/。"

The correct candidate personal name is "马诺哈尔·乔希" and not "马诺哈".

4.2 Location recognition

We collect 196 keywords such like "省,村,川,河,湖,角", when the system search these keywords in a string, it will collect some characters or words which maybe belong to a location with backward direction, and the candidate location can be inputted into location model to recognize. The approach is similar to the personal name recognition, the difference is its contextual, the contextual used for location is w_{i-2} w_{i-1} w_i w_{i+1} w_{i+2}, which always can be used as feature during location entity recognition.

We trained model based on the People's Daily.

We design some rules to help rectify wrong result, when a transliterated location name is lack of keyword like "省", it maybe recognized as a transliterated personal name. We collect some specific words list such as "奔赴,赴,圣地,故都" to correct the wrong personal name. If the current word is in the list, the following words are accepted as candidate location entity.

4.3 Organization recognition

Organization name recognition is very different from other kinds of entities. An organization is often composed of several characters or words, and its length is dynamitic. According to statistical result about People's Daily and MSR corpus, we decided the maximum length of an organization is 7 in a sentence.

We computed the probability of every word or character of an organization, and defined the probability threshold.

According to the different keyword, we designed sixteen classifiers; every classifier has its knowledge base, the different classifier can achieve organization recognition goal.

We computed the probability threshold (>0.02) of a candidate organization.

Combined the BIO-tagged method and the probability threshold, the organization can be recognized.

4.4 Combination of Knowledge from Various Sources

Human knowledge is very useful for NER. Knowledge from various sources can be incorporated into ME model, which are shown as follows.

1. Chinese family name list (including 925 items) and given names list (including 2453 items):

2. Transliterated character list (including 1398 items).

3. Location keyword list (including 607 items): If the word belongs to the list, 2~6 words before the salient word are accepted as candidate Location.

4. Abbreviated location like "京/Beijing", "津/Tianjin" name list. Moreover, on Microsoft corpus, the word "中" of "古今中外" is also labeled as location "中国/China".

5. Organization keyword list (including 875 items): If the current word is in organization keyword list, 2~6 words before keywords are accepted as the candidate Organization.

6. A location name dictionary. Some frequently used locations are included in the dictionary, like "美国/United States" and "新加坡/Singapore".

7. An organization name dictionary. Some frequently used organization names are included in the dictionary, like "国务院/State Council" and "联合国/United Nations".

8. Person name list: we collect some person names which come from the MSR train corpus. Moreover, the famous person name are included in the list such as "江泽民,李瑞环".

5 Evaluation result

We evaluated our word segmenter and named entity recognizer on the SIGHAN Microsoft Research Asia (MSRA) corpus in open track. The Table 1 is the official result of word segmentation by our system.

Corpus	OOV-Rate	OOV-Recall	IV Recall-rate	F measure
MSR	0.034	0.804	0.976	0.97
UPUC	0.087	0.593	0.957	0.911

Table 1 Official SIGHAN evaluation result for word segmentation in the open track

Table 2 shows the official result of entity recognition.

Type	R	P	F
Person	95.39%	96.71%	96.04%
Location	87.77%	93.06%	90.34%
Organization	87.68%	84.20%	85.90%

Table2 Official SIGHAN evaluation result for entity recognition in the open track

6 Conclusions

A probabilistic feature based ME model is used to Chinese word segmentation and named entity recognition tasks. Our word segmenter achieved the medium result in the open word segmentation track of MSRA corpus, while entity recognition achieved the top one performance.

Acknowledgement

The research work is supported by China Ministry Of Education funded project (MZ115-022): "Tools for Chinese and Minority Language Processing"

References

A L Berger. 1996. *A Maximum Entropy Approach to Natural Language Processing.* Computational Linguistic, 22 (1): 39- 71.

Yan Yintang, Zhou XiaoQiang. 2000.12 *Study of Segmentation Strategy on Ambiguous Phrases of Overlapping Type* Journal of The China Society For Scientific and Technical Information Vol. 19 , №6

Liang NanYuan. 1987 A *Written Chinese Segmentation system– CDWS*. Journal of Chinese Information Processing, Vol.2: 44-52

ZHANG Hua-ping and Liu Qun. 2004 *Automatic Recognition of Chinese Personal Name Based on Role Tagging*. CHINESE JOURNAL OF COMPUTERS Vol (27) pp 85-91.

Lv YaJuan, ZhaoTie-jun et al. 2001. *Leveled unknown Chinese Words resolution by dynamic programming*. Journal Information Processing, 15(1): 28-33.

Borthwick .A 1999. *Maximum Entropy Approach to Named Entity Recognition.* hD Dissertation.