Front. Electr. Electron. Eng. China (2006) 3: 273–276
DOI 10.1007/s11460-006-0037-9

RESEARCH ARTICLE

ZHANG Feng, FAN Xiao-zhong

Resolution of overlapping ambiguity strings based on maximum entropy model

Abstract The resolution of overlapping ambiguity strings (OAS) is studied based on the maximum entropy model. There are two model outputs, where either the first two characters form a word or the last two characters form a word. The features of the model include one word in context of OAS, the current OAS and word probability relation of two kinds of segmentation results. OAS in training text is found by the combination of the FMM and BMM segmentation method. After feature tagging they are used to train the maximum entropy model. The People Daily corpus of January 1998 is used in training and testing. Experimental results show a closed test precision of 98.64 % and an open test precision of 95.01 %. The open test precision is 3.76 % better compared with that of the precision of common word probability method.

Keywords Chinese information processing, Chinese automatic word segmentation, overlapping ambiguity strings, maximum entropy model

1 Introduction

Automatic word segmentation is a basis for Chinese information processing. The effect of word segmentation has direct influence on some studies such as machine translation, information retrieval and information extraction, etc.. Resolution of segmentating ambiguity strings is one of the problems of Chinese automatic word segmentation that has not been addressed well. It is also a reason that to influences the precision of word segmentation. There are two kinds of ambiguity strings: OAS and covering ambiguity strings (CAS) [1]. OAS is mainly a type of ambiguity string that accounts for more than 85 % of all ambiguity strings [2].

Translated from *Transactions of Beijing Institute of Technology*, 2005, 25(7): 590–593 (in Chinese)

ZHANG Feng (✉), FAN Xiao-zhong
Department of Computer Science and Engineering,
Beijing Institute of Technology, Beijing 100081, China
E-mail: feng.zhang@tom.com

Sun Mao-song et al. used the word probability method to solve the resolution of OAS with three Chinese characters [3]. Let ambiguity string be S, which includes three Chinese characters denoted as A, B, C. According to the two segmentation possibility, the occurrence probability of S is $P(AB, C) = P(AB)\,P(C)$ or $P(A, BC) = P(A)\,P(BC)$, where $P(AB, C)$ is the occurrence probability of S when the first two characters form a word, whereas $P(A, BC)$ is the occurrence probability of S when the last two characters form a word. Word probability method believes that the segmentation should be that which has better probability. This method is simple but its precision is not high. It can be viewed as the basic method of resolving of ambiguity strings. Li Rong et al. used the combination of SVM and k-NN [4] to study the resolution of ambiguity strings, which formalized the segmentation problem to a classification problem.

Because the segmentation problem of OAS has more than one overlapping strings that can be transformed to the problem of OAS, which has one overlapping string, they can be directly solved by using statistics rules [1]. The resolution of OAS with one overlapping string is mainly studied in this paper.

The resolution of OAS can be thought of as a binary classification problem, that is to say, either the first two characters form a word or the last two characters form a word. In recent years, the maximum entropy model has been widely used in language model and natural language processing. We use the maximum entropy model to solve the resolution of OAS in this paper.

2 Maximum entropy model for resolution of OAS

2.1 Common form of model

For the classification problem, let's set some samples (x, y), where x represents context, and y represents problem classes. We denote by $P(y|x)$ the probability that the model assigns to y in context x. The principle of maximum entropy states that the correct distribution $P(x, y)$ is that which maximizes entropy, or "uncertainty", subject to the constraints, which

represent "evidence", i.e., the facts known to the experimenter.

Mathematic measurement of the uncertainty of distribution $P(x, y)$ is entropy:

$$S(P) = -\sum_{x,y} P(x,y) lbP(x,y) \tag{1}$$

Mathematic measurement of the uncertainty of condition distribution $P(y|x)$ is condition entropy:

$$S_t(p) = -\sum_{x,y} \tilde{P}(x) P(y \mid x) lbP(y \mid x) \tag{2}$$

where P is the observed probability of x under some training sample. It can be calculated by maximum likelihood estimation.

Feature function is used to describe the known facts. Feature function is generally a binary valued function with the form:

$$f(x,y) = \begin{cases} 1 & \text{if}(x,y) \quad \text{satisfy} \quad \text{some} \quad \text{constraint } s \\ 0 & \text{else} \end{cases} \tag{3}$$

The constraints of the model can be expressed in the way that the expectation of the feature is equal to the expectation of the observed evidence:

$$E_P(f) = E_{\tilde{P}}(f) \tag{4}$$

where,

$$E_P f = \sum_{x,y} \tilde{P}(x) P(y \mid x) f(x,y) \tag{5}$$

$$E_{\tilde{P}} f = \sum_{x,y} \tilde{P}(x,y) f(x,y) \tag{6}$$

So given k features, the model set is

$$C = \{P \mid E_P f_i = E_{\tilde{P}} f_i, i = \{1, \cdots, k\}\} \tag{7}$$

Whereas the best model is P*, which has maximum entropy.

$$P^* = \arg\max_{P \in C} S(P) \tag{8}$$

It can be proved that P* meets the following form Ref. [8]:

$$P(y \mid x) = \frac{1}{Z(x)} \exp\left[\sum_{i=1}^{k} \lambda_i f_i(x,y)\right] \tag{9}$$

$$Z(x) = \sum_{y} \exp\left[\sum_{i=1}^{k} \lambda_i f_i(x,y)\right] \tag{10}$$

Form Eq. (9) is a general form of the maximum entropy model where k is the feature number and $Z(x)$ is the const variable for normalization to make sure $\sum_y p(y \mid x) = 1$. Parameter λ_i denotes degree of importance of feature f_i to the model.

2.2 Feature template

The advantage of the maximum entropy framework is that different features can be integrated in one model and has solid mathematical basis. Commonly, the OAS strings can get correct segmentation according to one word in its context. Therefore, we adapt one previous word pre(x) and one next word next(x) of ambiguity strings and the ambiguity strings self cur(x) as features. Besides, the probability relation of two segmentation possibilities is also taken as features. By the way. Because POS cannot be well tagged when segmentation is still not being completed, the model does not use POS information. For the segmentation of OAS, there are two problem classes: segmentation result when the first two characters form a word denoted as a, segmentation result when the last two characters form a word denoted as b. Table 1 shows the feature template of the maximum entropy model for resolution of OAS, where X represents context or current ambiguity string, T represents its segmentation class a or b.

Table 1 Feature templates for resolution of overlapping ambiguity strings

Features	Notes
pre (x) = X & y = T	Previous word and segmentation class
cur (x) = X & y = T	Current word and segmentation class
next (x) = X & y = T	Next word and segmentation class
P (AB, C) > P (A, BC) & y = T	The probability when the first two characters form a word is large than probability when the last two characters form a word, segmentation class
P (AB, C) < P (A, BC) & y = T	The probability when the first two characters form a word is less than probability when the last two characters form a word, segmentation class
P (AB, C) = P (A, BC) & y = T	The probability when the first two characters form a word is equal to probability when the last two characters form a word, segmentation class

For example, consider the sentence "一些生产和服务行业", where "和服务" is an OAS, whose segmentation class is b. The feature function is

$$f_1(x,y) = \begin{cases} 1 & \text{if} \quad \text{pre}(x) = \text{生产} \,\&\, y = b \\ 0 & \text{else} \end{cases}$$

$$f_2(x,y) = \begin{cases} 1 & \text{if} \quad \text{cur}(x) = \text{和服务} \,\&\, y = b \\ 0 & \text{else} \end{cases}$$

$$f_3(x,y) = \begin{cases} 1 & \text{if} \quad \text{next}(x) = \text{行业} \,\&\, y = b \\ 0 & \text{else} \end{cases}$$

$$f_3(x,y) = \begin{cases} 1 & \text{if} \quad P(AB,C) < P(A,BC) \,\&\, y = b \\ 0 & \text{else} \end{cases}$$

2.3 Feature selection

Feature selection and parameter estimation are two main tasks in the application of the maximum entropy model. Since not all features have contribution to the model, after the feature function is defined, occurrence features in the

observed data need to be selected. Furthermore, too many features can improve model training time. Count cutoff feature selection (CCFS) [5] and incremental feature selection (IFS) [6] are two commonly used methods in feature selection. Information gain of all candidate features will be calculated in IFS algorithm in every feature selection period, so it is time-consuming. In CCFS algorithm, given a threshold K, the model only considers those features whose occurrence times are larger than K. Though this method can not make sure that a minimum feature set is obtained, it is simple and effective in real applications. We adopt the CCFS algorithm in our experiment. The best threshold value is adjusted by testing precision. When the threshold value is 1, we got the best effect.

2.4 Parameter estimation

Another task of solving for the maximum entropy model is estimation of parameter λ. General estimation method includes generalized iterative scaling algorithm (GIS) proposed by Danroch and Ratcliff in 1972 and improved iterative scaling algorithm (IIS) proposed by Pietra et al. in 1995 [7−8]. Solution of model parameter in this paper is based on the IIS algorithm:

Input: feature functions $f_1, f_2, \cdots, f_n$; expectation distribution $P(x, y)$.

Output: the most optimized parameter λ, the most optimized model.

1) For all $i \in \{1, 2, \cdots, n\}$, set $\lambda_i = 0$.

2) For all $i \in \{1, 2, \cdots, n\}$, repeat the following step until all of λ_i is convergent

Get $\Delta\lambda_i$ in the following expression:

$$\sum_{x,y} \tilde{P}(x)P(y \mid x)f_i(x,y)\exp(\Delta\lambda_i f^{\#}(x,y)) = \tilde{P}(f_i) \qquad (11)$$

where $f^{\#}(x,y) \equiv \sum_{i=1}^{n} f_i(x,y)$.

Update λ_i according to $\lambda_i = \lambda_i + \Delta\lambda_i$.

3 Experiments

3.1 Data set

Data set used in experiments comes from the segmented and tagged corpus from the People Daily corpus of January 1998 of the Institute of Computational Linguistics in Peking University, which has almost 1 000 000 words. We extract overlapping ambiguity strings from this corpus and tag them. The extraction method is as follows: firstly, recover the segmented and tagged corpus to the original corpus. Secondly, segment the original strings into words by the combination of FMM and BMM segmentation method, and then extract strings from the different parts of two kinds of segmentation results and one word in context of them. These strings are just OAS candidates. After being corrected based on segmented corpus, they were tagged as output class *a* or *b*. Some samples of OAS are shown in Table 2. There are 8 213 instances totally. We randomly select 20 % for use in testing, and the rest for training.

Table 2 Some samples of training set

No.	Samples
1	pre(x) = sep cur(x) = 附图片 next(x) = sep $P(AB, C)$- $P(A, BC)$ < 0 *b*
2	pre(x) = 很 cur(x) = 不平凡 next(x) = 的 $P(AB, C)$ -$P(A, BC)$ <0 *b*
3	pre(x) = 全体 cur(x) = 中国人 next(x) = 的 P(AB, C)-P(A, BC) > 0 *a*
4	pre(x) = 还 cur(x) = 不安宁 next(x) = sep P(AB, C) -P(A, BC) <0 *b*
5	Pre(x) =克服 cur(x) = 和解决 next(x) = sep $P(AB,C)$ -$P(A, BC)$ <0 *b*
6	pre(x) = 总体 cur(x) = 要求和 next(x) = 各项 $P(AB, C)$ -$P(A, BC)$ >0 *a*
7	pre(x) = 这 cur(x) = 是非常 next(x) = 重要 P(AB, C) -P(A, BC) <0 *b*
8	pre(x) = sep cur(x) = 附图片 next(x) = sep P(AB, C) -P(A, BC) <0 *b*
9	pre(x) = sep cur(x) = 有机会 next(x) = 再次 P(AB, C) -P(A, BC) <0 *b*
10	pre(x) = 你们 cur(x) = 为首都 next(x) = 的 P(AB,C) -P(A,BC) <0 *b*
11	pre(x) = 只是 cur(x)= 一心想 next(x) = 着 P(AB,C) -P(A,BC) <0 *a*

Notes: sep represents some separator tags such as punctuation, number, English letter etc.

3.2 Results and analysis

The experimental environment is a Pentium 2.4 GHz PC with 512 MB of RAM and Windows 2000 operating system. Table 3 shows the experimental results. In addition, experimental results of word probability method are also given in this table.

Table 3 Experiment results

Training set	Test set	Closed test precision/%	Open test precision/%	Word probability/%
6 570	1 643	98.64	95.01	91.25

From Table 3, we can clearly see that very high precision has been obtained in the closed test by using the maximum entropy model. It proves that the model can simulate training data preferably. Besides, the open test precision is improved by 3.76 % compared with that of the precision of the common word probability method, which only uses the word probability relation of segmentation of OAS. In

contrast, the maximum entropy model can flexibly use more context features and predict preferably segmentations of new ambiguity strings.

4 Conclusions

The resolution of OAS is studied based on the maximum entropy model in this paper. The model takes the current OAS and word probability relation of two kinds of segmentation results as features. Context features can be obtained from OAS found by the combination of FMM and BMM segmentation method for real text, and word probability information comes from segmented and tagged corpus. Experimental results show that the maximum entropy model can present better results in the resolution of OAS compared with the word probability method.

References

1. Liu Kai-ying, Chinese Text Automatic Word Segmentation and Tagging, Beijing: The Commercial Press, 2000 (in Chinese)
2. Liang Nan-yuan. Written Chinese automatic word segmentation system-CDWS, Journal of Chinese Information Processing, 1987, 1(2): 44–52 (in Chinese)
3. Sun Mao-song, Zuo Zheng-ping, Huang Chang-ning, Algorithm for solving 3-character crossing ambiguities in Chinese word segmentation, Journal of Tsinghua University (Natural Science), 1999, 39(5): 101–103 (in Chinese)
4. Li Rong, Liu Shao-hui,Ye Shi-wei et al., A method of crossing ambiguities in Chinese word segmentation based on SVM and *k*-NN, Journal of Chinese Information Processing, 2001, 15(6): 13–18 (in Chinese)
5. Ratnaprkhi A., Maximum entropy Models for natural language ambiguity resolution, University of Pennsylvania, 1998
6. Berger A. L., Pietra S. A. D., Pietra V. J. D., A maximum entropy approach to natural language processing, Computational Linguistic, 1996, 22(1): 39–71
7. Darroch J. N., Ratcliff D., Generalized iterative scaling for log-linear models, The Annals of Mathematical Statistics, 1972, 43(5): 1470–1480
8. Pietra S. D., Pietra V. D., Lafferty J., Inducing features of random fields, IEEE Transactions on Pattern Analysis and Machine Intelligence, 1997, 19(4): 380–393