Unsupervised Training for Overlapping Ambiguity Resolution in Chinese Word Segmentation

Mu Li, Jianfeng Gao, Changning Huang
Microsoft Research, Asia
Beijing 100080, China
{t-muli,jfgao,cnhuang}@microsoft.com

Jianfeng Li
University of Science and Technology of China
Hefei, 230027, China
jackwind@mail.ustc.edu.cn

Abstract

This paper proposes an unsupervised training approach to resolving overlapping ambiguities in Chinese word segmentation. We present an ensemble of adapted Naïve Bayesian classifiers that can be trained using an unlabelled Chinese text corpus. These classifiers differ in that they use context words within windows of different sizes as features. The performance of our approach is evaluated on a manually annotated test set. Experimental results show that the proposed approach achieves an accuracy of 94.3%, rivaling the rule-based and supervised training methods.

1 Introduction

Resolving segmentation ambiguities is one of the fundamental tasks for Chinese word segmentation, and has received considerable attention in the research community. Word segmentation ambiguities can be roughly classified into two classes: overlapping ambiguity (OA), and combination ambiguity (CA). In this paper, we focus on the methods of resolving overlapping ambiguities.

Consider a Chinese character string ABC, if it can be segmented into two words either as AB/C or A/BC depending on different context, ABC is called an overlapping ambiguity string (OAS). For example, given a Chinese character string "各国有" (*ge4-guo2-you3*), it can be segmented as either "各 | 国有" (each state-owned) in Sentence (1) of Figure 1, or "各国 | 有" (every country has) in Sentence (2).

(1) 在 (in) | 各 (each) | 国有 (state-owned) | 企业 (enterprise) | 中 (middle)
(in each state-owned enterprise)

(2) 在 (in) | 人权 (human rights) | 问题 (problem) | 上 (on) | 各国 (every country) | 有 (have) | 共同点 (common ground)
(Regarding human rights, every country has some common ground)

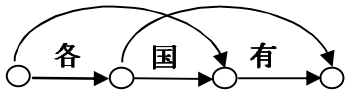

Figure 1. Overlapping ambiguities of Chinese character string "各国有"

Our method of resolving overlapping ambiguities contains two procedures. One is to construct an ensemble of Naïve Bayesian classifiers to resolve ambiguities. The other is an unsupervised method for training the Naïve Bayesian classifiers which compose the ensemble. The main issue of the unsupervised training is how to eliminate the negative impact of the OA errors in the training data. Our solution is to identify all OASs in the training data and replace them with a single special token. By doing so, we actually remove the portion of training data that are likely to contain OA errors. The classifiers are then trained on the processed training data.

Our approach is evaluated on a manually annotated test set with 5,759 overlapping segmentation ambiguities. Experimental results show that an accuracy of 94.3% is achieved.

This remainder of this paper is structured as follows: Section 2 reviews previous work. Section 3 defines overlapping ambiguous strings in Chinese. Section 4 describes the evaluation results. Section 5 presents our conclusion.

2 Previous Work

Previous methods of resolving overlapping ambiguities can be grouped into rule-based approaches and statistical approaches.

Maximum Matching (MM) based segmentation (Huang, 1997) can be regarded as the simplest rule-based approach, in which one starts from one end of the input sentence, greedily matches the longest word towards the other end, and repeats the process with the rest unmatched character sequences until the entire sentence is processed. If the process starts with the beginning of the sentence, it is called Forward Maximum Matching (FMM). If the process starts with the end of the sentence, it is called Backward Maximum Matching (BMM). Although it is widely used due to its simplicity, MM based segmentation performs poorly in real text.

Zheng and Liu (1997) use a set of manually generated rules, and reported an accuracy of 81% on an open test set. Swen and Yu (1999) presents a lexicon-based method. The basic idea is that for each entry in a lexicon, all possible ambiguity types are tagged; and for each ambiguity types, a solution strategy is used. They achieve an accuracy of 95%. Sun (1998) demonstrates that most of the overlapping ambiguities can be resolved without taking into account the context information. He then proposes a lexicalized rule-based approach. His experiments show that using the 4,600 most frequent rules, 51% coverage can be achieved in an open test set.

Statistical methods view the overlapping ambiguity resolution as a search or classification task. For example, Liu (1997) uses a word unigram language model, given all possible segmentations of a Chinese character sequence, to search the best segmentation with the highest probability. Similar approach can be traced back to Zhang (1991). But the method does not target to overlapping ambiguities. So the disambiguation results are not reported. Sun (1999) presents a hybrid method which incorporates empirical rules and statistical probabilities, and reports an overall accuracy of 92%. Li (2001) defines the word segmentation disambiguation as a binary classification problem. Li then uses Support Vector Machine (SVM) with mutual information between each Chinese character pair as a feature. The method achieves an accuracy of 92%. All the above methods utilize a supervised training procedure. However, a large manually labeled training set is not always available. To deal with the problem, unsupervised approaches have been proposed. For example, Sun (1997) detected word boundaries given an OAS using character-based statistical measures, such as mutual information and difference of t-test. He reported an accuracy of approximately 90%. In his approach, only the statistical information within 4 adjacent characters is exploited, and lack of word-level statistics may prevent the disambiguation performance from being further improved.

3 Ensemble of Naïve Bayesian Classifier for Overlapping Ambiguity Resolution

3.1 Problem Definition

We first give the formal definition of overlapping ambiguous string (OAS) and longest OAS.

An OAS is a Chinese character string *O* that satisfies the following two conditions:

a) There exist two segmentations Seg_1 and Seg_2 such that $\forall w_1 \in Seg_1, w_2 \in Seg_2$, where Chinese words w_1 and w_2 are different from either literal strings or positions;

b) $\exists w_1 \in Seg_1, w_2 \in Seg_2$, where w_1 and w_2 overlap.

The first condition ensures that there are ambiguous word boundaries (if more than one word segmentors are applied) in an OAS. In the example presented in section 1, the string "各国有" is an OAS but "各国有企业" is not because the word "企业" remains the same in both FMM and BMM segmentations of "各 | 国有 | 企业" and "各国 | 有 | 企业". The second condition indicates that the ambiguous word boundaries result from crossing brackets. As illustrated in Figure 1, words "各国" and "国有" form a crossing bracket.

The longest OAS is an OAS that is not a substring of any other OAS in a given sentence. For example, in the case "生活水平" (*sheng1-huo2-shui3-ping2*, living standard), both "生活水" and "生活水平" are OASs, but only "生活水平" is the longest OAS because "生活水" is a substring of "生活水平". In this paper, we only consider the longest OAS because both left and right boundaries of the longest OAS are determined.

Furthermore, we constrain our search space within the FMM segmentation O_f and BMM segmentation O_b of a given longest OAS. According to Huang (1997), two important properties of OAS has been identified: (1) if the FMM segmentation is the same as its BMM segmentation ($O_f = O_b$), for example "搜索引擎" (*sou1-suo3-yin3-qing2*, Search Engine), the probability that the MM segmentation is correct is 99%; Otherwise, (2) if the FMM segmentation differs from its BMM segmentation ($O_f \neq O_b$), for example "各国有", the probability that at least one of the MM segmentation is correct is also 99%. So such a strategy will not lower the coverage of our approach.

Therefore, the overlapping ambiguity resolution can be formulized as a binary classification problem as follows:

Given a longest OAS *O* and its context feature set *C*, let *G*(*Seg*, *C*) be a score function of *Seg* for $seg \in \{O_f, O_b\}$, the overlapping ambiguity resolution task is to make the binary decision:

$$seg = \begin{cases} O_f & G(O_f, C) > G(O_b, C) \\ O_b & G(O_f, C) < G(O_b, C) \end{cases} \quad (1)$$

Note that $O_f = O_b$ means that both FMM and BMM arrive at the same result. The classification process can then be stated as:

a) If $O_f = O_b$, then choose either segmentation result since they are same;

b) Otherwise, choose the one with the higher score *G* according to Equation (1).

For example, in the example of "搜索引擎", if $O_f = O_b$ = "搜索 | 引擎", then "搜索 | 引擎" is selected as the answer. In another example of "各国有" in sentence (1) of Figure 1, O_f = "各国 | 有", O_b = "各 | 国有". Assume that *C* = {在, 企业}, i.e., we used a context window of size 3; then the segmentation "各国 | 有" is selected if *G*("各国|有",{在,企业}) > *G*("各 | 国有",{在, 企业}) , otherwise "各 | 国有" is selected.

3.2 Naïve Bayesian Classifier for Overlapping Ambiguity Resolution

Last section formulates the overlapping ambiguity resolution of an OAS *O* as the binary classification between O_f and O_b. This section describes the use of the adapted Naïve Bayesian Classifier (NBC) (Duda and Hart, 1973) to address problem. Here, we use the words around *O* within a window as features, with $w_{-m}\ldots w_{-1}$ denoting *m* words on the left of the *O* and $w_1\ldots w_n$ denoting *n* words on the right of the *O*. Naïve Bayesian Classifier assumes that all the feature variables are conditionally independent. So the joint probability of observing a set of context features $C = \{w_{-m}\ldots w_{-1}, w_1\ldots w_n\}$ of a segmentation *Seg* (O_f or O_b) of *O* is as follows:

$$p(w_{-m}, w_{-1}, w_1, \ldots w_n, Seg) = p(Seg) \prod_{i=-m,\ldots-1,1,\ldots n} p(w_i \mid Seg) \quad (2)$$

Assume that Equation (2) is the score function in Equation (1) *G*, we then have two parameters to be estimated: $p(Seg)$ and $p(w_i|Seg)$. Since we do not have enough labeled training data, we then resort to the redundancy property of natural language. Due to the fact that the OAS occupies only in a very small portion of the entire Chinese text, it is feasible to estimate the word co-occurrence probabilities from the portion of corpus that contains no overlapping ambiguities. Consider an OAS 信心地 (*xin4-xin1-de*, confidently). The correct segmentation would be "信心 | 地", if 充满 (*cong1-man3*, full of) were its context word. We note that 充满 appears as the left context word of 信心 in both strings 充满信心地 and 充满信心和勇气 (勇气, *yong3-qi4*, courage). While the former string contains an OAS, the latter does not. We then remove all OAS from the training data, and estimate the parameters using the training data that do not contain OAS. In experiments, we replace all longest OAS that has $O_f \neq O_b$ with a special token `[GAP]`. Below, we refer to the processed corpus as tokenized corpus.

Note that *Seg* is either the FMM or the BMM segmentation of *O*, and all OASs (including *Seg*) have been removed from the tokenized corpus, thus there are no statistical information available to estimate $p(Seg)$ and $p(w_{-m}\ldots w_{-1}, w_1\ldots w_n|Seg)$ based on the Maximum Likelihood Estimation (MLE) principle. To estimate them, we introduce the following two assumptions.

1) Since the unigram probability of each word *w* can be estimated from the training data, for a given segmentation $Seg = w_{s1}\ldots w_{sk}$, we assume that each word *w* of *Seg* is generated independently. The probability $p(Seg)$ is approxi-

mated by the production of the word unigram probabilities:

$$p(Seg) = \prod_{w_i \in Seg} p(w_i) \quad (3)$$

2) We also assume that left and right context word sequences are only conditioned on the leftmost and rightmost words of *Seg*, respectively.

$$\begin{aligned} & p(w_{-m}...w_{-1}, w_1...w_n \mid Seg) \\ & = p(w_{-m}...w_{-1} \mid w_{s1}) p(w_1...w_n \mid w_{sk}) \\ & = \frac{p(w_{-m}...w_{-1}, w_{s1}) p(w_{sk}, w_1...w_n)}{p(w_{s1}) p(w_{sk})} \end{aligned} \quad (4)$$

where the word sequence probabilities $P(w_{-m}, \ldots, w_{-1}, w_{s1})$ and $P(w_{sk}, w_1, \ldots, w_n)$ are decomposed as productions of trigram probabilities. We used a statistical language model toolkit described in (Gao et al, 2002) to build trigram models based on the tokenized corpus.

Although the final language model is trained based on a tokenized corpus, the approach can be regarded as an unsupervised one from the view of the entire training process: the tokenized corpus is automatically generated by an MM based segmentation tool from the raw corpus input with neither human interaction nor manually labeled data required.

3.3 Ensemble of Classifiers and Majority Vote

Given different window sizes, we can obtain different classifiers. We then combine them to achieve better results using the so-called ensemble learning (Peterson 2000). Let NBC(*l*, *r*) denote the classifier with left window size *l* and right window size *r*. Given the maximal window size of 2, we then have 9 classifiers, as shown in Table 1.

	$L = 0$	$l = 1$	$l = 2$
$r = 0$	NBC(0, 0)	NBC(1, 0)	NBC(2,0)
$r = 1$	NBC(0,1)	NBC(1,1)	NBC(2,1)
$r = 2$	NBC(0,2)	NBC(1,2)	NBC(2,2)

Table 1. Bayesian classifiers in the ensemble

The ensemble learning suggests that the ensemble classification results are based on the majority vote of these classifiers: The segmentation that is selected by most classifiers is chosen.

4 Experiments and Discussions

4.1 Settings

We evaluate our approach using a manually annotated test set, which was selected randomly from People's Daily news articles of year 1997, containing approximate 460,000 Chinese characters, or 247,000 words. In the test set, 5759 longest OAS are identified. Our lexicon contains 93,700 entries.

4.2 OAS Distribution

We first investigate the distribution of different types of OAS in the test set. In our approach, the performance upper bound (i.e. oracle accuracy) cannot achieve 100% because not all the OASs' correct segmentations can be generated by FMM and BMM segmentation. So it is very useful to know to what extent our approach can deal with the problem.

The results are shown in Table 2. We denote the entire OAS data set as *C*, and divide it into two subsets *A* and *B* according to the type of OAS. It can be seen from the table that in data set *A* (O_f=O_b), the accuracy of MM segmentation achieves 98.8% accuracy. Meanwhile, in data set *B* ($O_f \neq O_b$) the oracle recall of candidates proposed by FMM and BMM is 95.7% (97.2% in the entire data set *C*). The statistics are very close to those reported in Huang (1997).

A $OAS_{Of = Ob}$ 2763 47.98%	$O_f = O_b$ = COR 2731 47.42%
	$O_f = O_b \neq$ COR 32 0.56%
B $OAS_{Of \neq Ob}$ 2996 52.02%	O_f = COR $\vee$ O_b = COR 2866 49.77%
	$O_f \neq$ COR $\wedge$ $O_b \neq$ COR 130 2.26%

Table 2. Distribution of OAS in the test set

Here are some examples for the overlapping ambiguities that cannot be covered by our approach. For errors resulting from $O_f = O_b \neq$ COR, a typical example in the literature is 结合成分子时 (jie2-he2-cheng2-fen1-zi3-shi2, 结合 | 成 | 分子 |

时). For errors caused by $O_f \neq O_b$ and $O_f \neq \text{COR} \wedge O_b \neq \text{COR}$, 出现在世纪(之初) (chu1-xian4-zai4-shi4-ji4, 出现 | 在 | 世纪) serves as a good example. These two types of errors are usually composed of several words and need a much more complicated search process to determine the final correct output. Since the number of such errors is very small, they are not target of our approach in this paper.

4.3 Experimental Results of Ensemble of Naïve Bayesian Classifiers

The classifiers are trained from the People's Daily news articles of year 2000, which contain over 24 million characters. The training data is tokenized. That is, all OAS with $O_f \neq O_b$ are replaced with the token [GAP]. After tokenization, there are 16,078,000 tokens in the training data in which 203,329 are [GAP], which is 1.26% of the entire training data set. Then a word trigram language model is constructed on the tokenized corpus, and each Bayesian classifier is built given the language model.

	$l = 0$	$l = 1$	$l = 2$
$r = 0$	88.73%	88.85%	88.95%
$r = 1$	89.09%	89.39%	89.39%
$r = 2$	88.95%	89.39%	89.35%

Table 3. Accuracy of each individual classifier

Table 3 shows the accuracy of each classifier on data set *B*. The performance of the ensemble based on majority vote is 89.79% on data set *B*, and the overall accuracy on *C* is 94.13%. The ensemble consistently outperforms any of its members. Classifiers with both left and right context features perform better than the others because they are capable to segment some of the context sensitive OAS. For example, contextual information is necessary to segment the OAS "看台上"(*kan4-tai2-shang4*, on the stand) correctly in both following sentences:

你 | 看 | 台上 | 那个 | 演员
(Look at the performer in the stage)
站 | 在 | 最高 | 一 | 层 | 看台 | 上
(Stand in the highest stand)

Both Peterson (2000) and Brill (1998) found that the ultimate success of an ensemble depends on the assumption that classifiers to be combined make complementary errors. We investigate this assumption in our experiments, and estimate the oracle accuracy of our approach. Result shows that only 6.0% (180 out of 2996) of the OAS in data set *B* that is classified incorrectly by all the 9 classifiers. In addition, we can see from Table 2, that 130 instances of these 180 errors are impossible to be correct because neither O_f nor O_b is the correct segmentation. Therefore, the oracle accuracy of the ensemble is 94.0%, which is very close to 95.7%, the theoretical upper bound of our approach in data set *B* described in Section 4.2. However, our majority vote based ensemble only achieves accuracy close to 90%. This analysis thus suggests that further improves can be made by using more powerful ensemble strategies.

4.4 Lexicalized Rule Based OAS Disambiguation

We also conduct a series of experiments to evaluate the performance of a widely used lexicalized rule-based OAS disambiguation approach. As reported by Sun (1998) and Li (2001), over 90% of the OAS can be disambiguated in a context-free way. Therefore, simply collecting large amount of correctly segmented OAS whose segmentation is independent of its context would yield pretty good performance.

We first collected 730,000 OAS with $O_f \neq O_b$ from 20 years' People's Daily corpus which contains about 650 million characters. Then approximately 47,000 most frequently occurred OASs were extracted. For each of the extracted OAS, 20 sentences that contain it were randomly selected from the corpus, and the correct segmentation is manually labeled. 41,000 lexicalized disambiguation rules were finally extracted from the labeled data, whose either MM segmentation (O_f or O_b) gains absolute majority, over 95% in our experiment. The rule set covers approximately 80% occurrences of all the OASs in the training set, which is very close to that reported in Sun (1998). Here is a sample rule extracted: 信心地 => 信心 | 地. It means that among the 20 sentences that contain the character sequence "信心地", at least 19 of them are segmented as "信心 | 地".

The performance of the lexicalized rule-based approach is shown in Table 4, where for compari-

son we also include the performance of using only FMM or BMM segmentation algorithm.

	Accuracy	
	Data set B	Data set C
FMM	49.44%	73.12%
BMM	46.31%	71.51%
Rule + FMM	83.10%	90.65%
Rule + BMM	84.43%	91.33%
NBC(0, 0)	88.73%	93.70%
Ensemble	89.79%	94.13%

Table 4. Performance comparison

In Table 4, Rule + FMM means if there is no rule applicable to an OAS, FMM segmentation will be used. Similarly, Rule + BMM means that BMM segmentation will be used as backup. We can see in Table 4 that rule-based systems outperform their FMM and BMM counterparts significantly, but do not perform as well as our method, even when no context feature is used. This is because that the rules can only cover about 76% of the OASs in the test set with precision 95%, and FMM or BMM performs poorly on the rest of the OASs. Although the precision of these lexicalized rules is high, the room for further improvements is limited. For example, to achieve a higher coverage, say 90%, much more manually labeled training data (i.e. 81,000 OAS) are needed.

5 Conclusion and Future work

Our contributions are two-fold. First, we propose an approach based on an ensemble of adapted naïve Bayesian classifiers to resolving overlapping ambiguities in Chinese word segmentation. Second, we present an unsupervised training method of constructing these Bayesian classifiers on an unlabeled training corpus. It thus opens up the possibility for adjusting this approach to a large variety of applications. We perform evaluations using a manually annotated test set. Results show that our approach outperforms a lexicalized rule-based system. Future work includes investigation on how to construct more powerful classifier for further improvements. One promising way is combining our approach with Sun's (1997), with a core set of context free OASs manually labeled to improve accuracy.

Acknowledgements

We would like to thank Wenfeng Yang and Xiaodan Zhu for helpful discussions on this project and Wenfeng's excellent work on the lexicalized disambiguation rule set construction.

References

Eric Brill and Wu Jun. 1998. Classifier combination for improved lexical disambiguation. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics*, Montreal. CA.

Richard Duda and Peter Hart. 1973. *Pattern Classification and Scene Analysis*. Wiley, New York, NY.

Jianfeng Gao and Joshua Goodman, Li Mingjing, Lee Kai-Fu. 2002. Toward a unified approach to statistical language modeling for Chinese. *ACM Transactions on Asian Language Information Processing*, 1 (1):3-33.

Changning Huang. 1997. Segmentation problem in Chinese Processing. (In Chinese) *Applied Linguistics*. 1:72-78.

Rong Li, Shaohui Liu, Shiwei Ye and Zhongzhi Shi. 2001. A Method of Crossing Ambiguities in Chinese word Segmentation Based on SVM and k-NN. (In Chinese) *Journal of Chinese Information Processing*, 15(6): 13-18.

Ting Liu, Kaizhu Wang and Xinghai Jiang. 1997. The Maximum Probability Segmentation Algorithm of Ambiguous Character Strings. (In Chinese) *Language Engineering*. Tsinghua University Press. pp.182-187.

Ted Pedersen. 2000. A Simple Approach to Building Ensembles of Naive Bayesian Classifiers for Word Sense Disambiguation. In *Proceedings of the First Annual Meeting of the North American Chapter of the Association for Computational Linguistics*. Seattle, WA. pp. 63-69

Maosong Sun, Changning Huang and Benjamin K. Tsou. 1997. *Using Character Bigram for Ambiguity Resolution In Chinese Words Segmentation*. (In Chinese) Computer Research and Development, 34(5): 332-339

Maosong Sun and Zhengping Zuo. 1998. Overlapping ambiguity in Chinese Text. (In Chinese) *Quantitative and Computational Studies on the Chinese Language*, HK, pp. 323-338

Maosong Sun, Zhengping Zuo and Changning Huang. 1999. Algorithm for solving 3-character crossing ambiguities in Chinese word segmentation. (In Chi-

nese) *Journal of Tsinghua University Science and Technology*, 39(5).

Bing Swen and Shiwen Yu. 1999. A Graded Approach for the Efficient Resolution of Chinese Word Segmentation Ambiguities. In *Proceedings of 5th Natural Language Processing Pacific Rim Symposium*. pp. 19-24

Junsheng Zhang, Zhida Chen and Shunde Chen. 1991. Constraint Satisfaction and Probabilistic Chinese Segmentation. (In Chinese) In *Proceedings of ROCLING IV* (*R.O.C. Computational Linguistics Conference*). pp. 147-165

Jiaheng Zheng and Kaiying Liu. 1997. The Research of Ambiguity Word – Segmentation Technique for the Chinese Text. (In Chinese) *Language Engineering*, Tsinghua University Press, pp. 201-206