New Word Detection Algorithm for Chinese Based on Extraction of Local Context Information

Hua-Lin Zeng, Chang-Le Zhou, Xiao-Dong Shi, Tang-Qiu Li, Chang Su
Department of Cognitive Science，Xiamen University, Xiamen 361005，China
{hlzeng, dozero, mandel, tqli, suchang }@xmu.edu.cn

Abstract

Chinese segmentation is an important issue in Chinese text processing. The traditional segmentation methods those depend on an existing dictionary suffer the drawbacks when encounter unknown words. The paper proposed a segmenting algorithm for Chinese based on extracting local context information. It added the context information of the testing text into the local PPM statistical model so as to guide the detection of new words. The algorithm focusing on the process of online segmentation and new word detection achieves a good effect in the close or opening test, and outperforms some well-known Chinese segmentation system to a certain extent.

1. Introduction

Almost all techniques for statistical language processing are word based. This kind of language modeling works very well for western languages, where the words are well defined, but it is quite difficult to apply for Chinese. Chinese language is based on characters. There are no spaces between characters and word boundaries are not explicitly marked. Because words in Chinese are actually not well defined, so there does not exist a commonly accepted lexicon. Furthermore, the segmentation of a sentence into a string of words is not unique. Out of vocabulary (OOV) problem is especially serious, which would have more substantial impact on the performance of word segmentation than ambiguous segmentation [1]. For example, approximately 30% of OOV words in the SIGHAN's PK corpus are new words of this type [2].

978-1-4244-2197-8/08/$25.00 ©2008 IEEE

In the process of Chinese words segmentation, the general training method is off-line. Training corpora are prepared in advanced, so there is a high correlation between the result of segmentation and the type of training corpora. When the training corpora are different types from the testing text, the result may be even worse. As to the words in the lexicon and some OOV which could be processed with morphology rules, rule-based method works well, but this kind of method can not deal with new words, or words that no rules could be based or rules haven't been collected yet.

Making a comprehensive view on the segmenting text, most OOV appear many times in the context which could be called the multi-words units (MWU) or tokens, so we can extract and collect these MWU from the context, and use them to improve the segmentation result. One easy way is to regard these MWU as OOV. This method would complicate the segmentation rules, and also destroy the integrality of the rules. This paper presents an algorithm which adds the local context information to the statistical language model to improve the effect of the words segmentation. Good performance was obtained in new word identification by our experiment.

2. Improved PPM Word Segmentation Model

Cleary and Witten [3] presented a language model – PPM, prediction by partial matching, which was developed in the field of text compression in 1984. Though many other compression techniques exist, PPM has become a benchmark in the compression community and has been widely used in language processing tasks, such as character level language

modeling. It is an n-gram approach that uses finite-context models of characters, where the previous few characters predict the upcoming one. This characteristic focusing on character level connects PPM model with HMM model in Markov Properties of Limited Horizon. The PPM text compression model can be used together with HMM to identify tokens in text [4]. PPM model is an adaptive model that both decoder and encoder maintain the same statistical model- not by communicating the models directly, but by updating them in precisely the same way.

PPM model presents escape method on data sparseness problem. Consider the case where the context has occurred but never followed by the upcoming character which is called the zero-frequency situation, the encoder and decoder will escape at the same time with some special probability for smoothing the data. PPM uses an escape event to drop the model down to a lower order.

Define $p(e_i \mid e_{i-n+1}^{i-1})$ to be the probability of the preceding context e_{i-n+1}^{i-1} predicting the probability of the upcoming symbol e_i. Several different methods have been proposed on how to estimate $\alpha(e_i \mid e_{i-n+1}^{i-1})$ and $\gamma(e_{i-n+1}^{i-1})$, which are respectively named PPMA, PPMB,PPMC and PPMD.

$$p(e_i \mid e_{i-n+1}^{i-1}) = \begin{cases} \alpha(e_i \mid e_{i-n+1}^{i-1}) & if \quad c(e_i \mid e_{i-n+1}^{i-1}) > 0 \\ \gamma(e_{i-n+1}^{i-1}) p(e_i \mid e_{i-n+2}^{i-1}) & if \quad c(e_i \mid e_{i-n+1}^{i-1}) = 0 \end{cases}$$

There are two types of word segmentation. One is separation, the other is combination. Separation means the text to be segmented is looked as a character sequence union, which need to be separated in the correct positions. There are some typical algorithms such as Max-matched and Min-matched method. Combination means that the text to be segmented is looked as a character sequence individual, each character is independent and there exists a potential space between each two character, so the segmentation task is to combine the most likely characters as words. Some probable methods such as mutual information, maximum entropy, and conditional random fields are some typical ones.

Using PPM to segment the text in our experiment we adopt combination model. Regarding that between every two characters, there is a binary probability function, so when the function takes 1, the space exists legally and separation should be filled at this position; when the function takes 0, the space doesn't exist and there is no separation. This process may be presented by HMM model also.

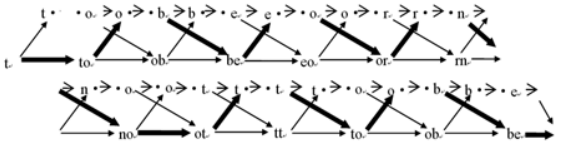

Figure 1. 1st-order PPM modeling of string *tobeornottobe*

In figure 1, each node is a 1st order PPM composed with one context including one processing character and the character looking forward. We could encode probability of each state through PPM. Bold line presents correct search path, means the correct segmentation sequence of this string, and this is the maximum value we found of the observation sequence. W.J.Teahan etc [5] exploits a PPM-based algorithm, which brings traditional PPM model into Chinese word segmentation task by searching one path that using PPM model to compress the text and minimize the bits of compressed code, which achieved a fine performance. In our experiment, we improve it by only using PPM model to calculate probability of each state but not using it to compress the text actually, that is using the probability estimating formula to find a path to maximize the accumulative value of segmented sequence through PPM modeling. Define $\Pr(S)$ as the value of segmented sequence, S_i as one node in the search map, W_i as the processing character and W_{i-n} as the context looking forward n characters.

$$\Pr(S) = \prod_i \Pr(S_i) = \prod_i \Pr(W_i \mid W_{i-n})$$

3. Effect of Local Context Information

Given two examples to analyze how the local context information of segmenting text works on the segmentation result.

Example 1: Help to discovering of new words.

"在党的十七大胜利闭幕后不久，这样的短信息就出现在广西扶绥县党员的手机屏幕上。通过'短信党课'教育平台，扶绥县向1．4万多名党员传达了党的十七大精神，使党的十七大精神迅速传遍千家万户。"

Note that in this short paragraph, OOV word 十七大 appears three times. If using this information appropriately, we could find out some new words appearing several times in the text to improve the problem which are difficult for off-line methods.

Example 2: Help to disambiguation of combinatorial ambiguous phrase.

“王芳超生前使用过的物品”

Segmentation 1：王芳超/nr 生前/v 使用/v 过/v 的/u 物品/n

Segmentation 2：王芳/nr 超生/v 前/f 使用/v 过/v 的/u 物品/n

The two segmentations are totally different on semantics, but they are both proper on syntax, which can not be judged through rule-based or statistical-based method only, but if MWU 王芳超 could be found several times in the context, segmentation 2 is the correct result.

4. Extracting Local Context Information

4.1 Extracting Local Context Information

Extracting information from the local context means extracting MWU which have the highest reliability. In our experiment, we use Suffix Array [6] to extract candidate entries, improving the efficiency of the search process. There follows our algorithm on extracting:

1. Construct suffix array ordering by the left and right affix of words to build Left-index and Right-index array.
2. Compare the left suffix array and right suffix array, find n-gram token both appearing in two array and add them to c-list. Sort the tokens in c-list and extract the most frequent n-gram token as candidate entry.
3. Consider the candidate entries to find if they can compose a word and not overlay ambiguity or wrong extracting because of high frequency co-occurrences, then delete the wrong entries by segmentation program.

3.1 Delete Stop Words.

Some single-character empty word which have low probability to compose a word, such as 的了们是个在说和是为有到 , should be deleted and not be added into the local probability model if they appeared in the candidate entries.

3.2 Delete the entries covering traditional words.

For finding the highest frequency suffix, many co-occurrences would be chosen and stored, such as 建设社会主义、法律法规、共同理想 etc., or some entries covering lexicon word such as tokens 们发现了 and so on. These tokens include some lexicon word, so we carry out a chart-based search on the tokens having the highest frequency. If the token covers a lexicon word, it should not be the final candidate entry. Furthermore, if the token could be segmented by PPM model, it should not be counted.

4. Add the final entries into the segmenting statistical model as the reference of local context information on the segmentation.

In the experiment, we find that if we extracting the longest co-occurrence units using the segmenting text directly, it doesn’t contribute to the result but have severe side-effect. In the survey of Sinica Corpus, 4572 entries are wrongly divided into shorter one-character words of all 4632 OOV [7]. So one-character word means the possibility of OOV. A supplement method is to survey the one-character lexicon word. For that reason, during extracting n-gram MWU, we only extract the fragments of segmentations which increase the segmentation performance greatly.

4.2 Training Context PPM Model

The PPM model training by the training corpora is called the global model, and training by the entries of extracting the local context from the testing corpora and training locally is called the local model.

MLE is applied as our training method to count all n-order pairs of the state <context, nextchar>. In the experiment, we set different weight to the training data, not adding 1 only, this method decreases wrong segmentations caused by the manual labeling.

The training data we chose are as below:

1. Tagged corpus of People Daily of Jan 1998 from Institute of Computational Linguistics (ICL) at Peking University, including 1,839,490 pieces of words, 8.09MB.
2. An electro-dictionary form SEGTAG of Xiamen University, including 88,694 pieces of words.

According to two training data, the experiment uses two different methods.

Global training means training with the sentences. For example, input the sentence:

“迈向/v 充满/v 希望/n 的/uj 新/a 世纪/n”

After the preprocessing, we get the continuous Chinese character sequence:

“ 迈向 充满 希望 的 新 世纪 ”

Local training means training with the lexicons in the dictionary.

After the preprocessing, we get:

“ 迈向 ”, “ 充满 ”, “ 希望 ”, “ 的 ”, “ 新 ”, “ 世纪 ”

We totally get six words and train with them. This method is training the dictionary actually.

Two methods have their own characteristics: the global one considers the boundary of lexicon words and have a better performance on the occurrence of words. But the training model is huge and the data is sparse, so the efficiency of system is slow. The local one does well in the structure of one word, but ignores the occurrence of words. The training model is small and faster than the global one.

Global and local model play a different role in the segmentation job. The global model is used to segment the lexicon words, and the local model is used to find the new words. Some OOV have some rules of construction which could be found out through the training of global model; but to some new words or some words never appearing in the training corpus, we could find the corresponding information through the local model.

5. Evaluation

The performance of word segmentation is measured through test precision (P), test recall (R), F score (which is defined as 2PR/(P+R)), the OOV rate for the test corpus.

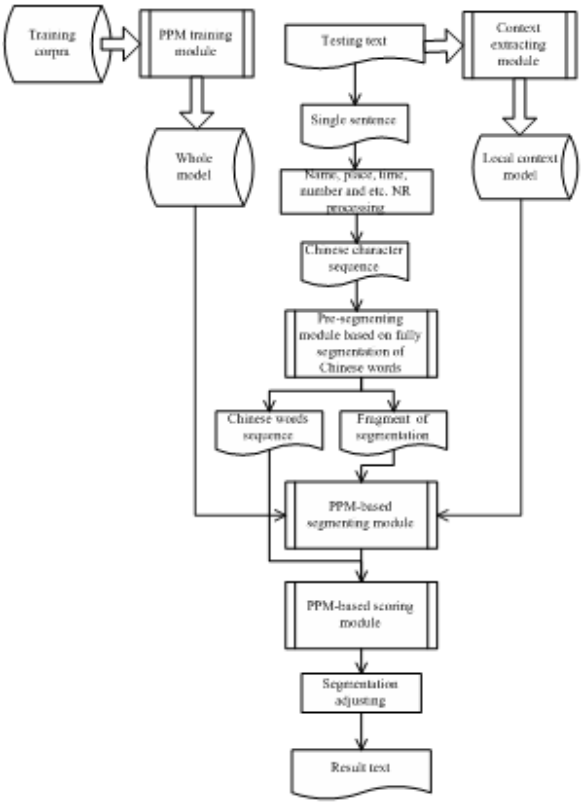

Figure 2. Flow chart of XMSEG

5.1 System performance on Chinese words segmentation

We evaluated the improved PPM-based word segmentation system (henceforth XMSEG) using the training corpora – standards used in SIGHAN's Second International Chinese Word Segmentation Bakeoff (see Sproat and Emperson (2005) for details). The training and test corpora are annotated manually, where there is only one allowable word segmentation for each sentence. The systems for compare are SEGTAG 1.17 from Xiamen University, ICTCLAS from ICT, and S-MSRseg V1 from MSRA.

From table 1, we could see that using ICTCLAS testing corpora – People's Daily of Jan 1998, except ICTCLAS, our system achieves a best result than other two, and from table 2, using MSRseg testing corpora – 1000 sentences downloaded from MSRA, except S-MSRseg, our system also gets best result. By examining the segmentation results we could see our system finds out many OOV that other system couldn't find, this shows the good segmenting performance in the circumstance when the testing corpora and training corpora are not the same type.

Table 1. Performance of Chinese segmentation using ICTCLAS testing corpora

Testing corpora	ICTCLAS testing corpora		
	F Score	Recall	Precision
XMSEG	97.0141%	96.3289%	97.7093%
SEGTAG	96.2915%	96.2915%	97.0998%
S-MSRseg	85.7161%	88.5537%	83.0548%
ICTCLAS	98.9953%	98.8658%	99.1252%

Table 2. Performance of Chinese segmentation using MSRseg testing corpora

Testing corpora	MSRseg testing corpora		
	F Score	Recall	Precision
XMSEG	93.6682%	93.4270%	93.9106%
SEGTAG	92.7387%	93.3661%	92.1197%
S-MSRseg	92.9730%	93.8109%	92.1499%
ICTCLAS	98.4657%	98.2491%	98.6833%

5.2 Performance of New Words Detection

In this experiment, we download the testing corpora on the testing day from website – www.sina.com.cn randomly which contain 32 pieces of news, totally 87KB. The testing corpora involve science, entertainment, society culture, physical culture and 3 pieces of serial novel, which contain 84,673 Chinese characters and are the latest corpora which have no relations with the training corpora of all four testing system. It is an open test. .This test takes manually evaluating. We classify the OOV into three parts: Chinese Person Name, Foreign Person Name and New Words (including new words of internet, place name and location name). The Result again show the advantage of our method in new word detection.

There are some examples that OOV recognized by XMSEG:

Foreign Person Name: 俵义文, 宫崎, 潮匡人市, 西村幸佑, 柯林斯, etc.

Chinese Person Name: 蓝建中, 张谕, 冯媛庆,王智新, 郑洋 etc.

New Words: 雅虎, 国美, 永乐, 环评, 鹰谷, 南南合作 etc.

Table 3. Comparison of Performance in Detection of New Words

Precision	XMSEG	ICTCLAS	SEGTAG	MSRseg
Foreign Person Name	88.6364%	69.6970%	70.4545%	68.9394%
Chinese Person Name	87.3016%	30.1587%	42.8571%	39.6825%
New Words	96.5116%	34.8837%	67.4419%	46.5116%

6. Conclusion

This paper takes a in-depth investigation in text compression algorithm – PPM, discusses the role of PPM playing in the Chinese words segmentation task, models the Chinese segmentation by HMM model, and finally presents an on-line Chinese words segmenting algorithm based on PPM model. The algorithm implements easily and takes a good performance also when there is a lack of the training corpora. But the disadvantage of our algorithm is the high complexity and long executing time. The improvement algorithm takes a preprocessing segmenting which achieves a great improvement of segmenting efficiency. And based on the characteristic of the local appearance of Chinese character, we extract the context information using Suffix Array, not simply extracting the MWU appearing many times in the local context but adding this information into the global statistical model to conduct the process of OOV in the testing text. This algorithm combines the complete segmenting preprocessing, processes on the segmenting fragments, and extracts useful information from them. The experiment result proves that this algorithm is correct and effective.

Acknowledgments

The project was supported by the Foundation of Fujian Province of China under Grant No. 2006H0038 and the Foundation of Fujian Province of China under Grant No. 2008F3105.

References

[1] Sun, Maosong, and Zou Jiayan, “A critical appraisal of the research on Chinese word segmentation”, *Contemporary Linguistics*, Jan 2001.

[2] Gao, Jianfeng, Andi Wu, Mu Li, Chang-Ning Huang, Hongqiao Li, Xinsong Xia, and Haowei Qin, “Adaptive Chinese word segmentation”. *ACL2004*. July 21-26.

[3] Cleary, J. G., and Witten, I. H., “Data compression using adaptive coding and partial string matching”, *IEEE Transactions on Communications*, 1984, 32(4): pp. 396–402.

[4] Yingying Wen, Ian H. Witten, and Dianhui Wang, “Token Identification Using HMM and PPM Models”, *AI 2003*, LNAI 2903, 2003:pp. 173–185.

[5] W. J. Teahan, Y. Wen, R. McNab, and I. H. Witten., “A Compression-based Algorithm for Chinese Word Segmentation”, *Computational Linguistics*, 2000, 26(3): pp. 375–393.

[6] Chien, and Lee-Feng, “PAT-tree-based keyword extraction for Chinese information retrieval”, *IGIR97*, 2007:pp.27-31.

[7] Zhiyong Luo, and Rou Song, “An Integrated Method for Chinese Unknown Word Extraction”, *Proceedings of 3rd ACL SIGHAN Workshop*, 2004: pp. 148-15.