Proceedings of the IASTED International Conference
ARTIFICIAL INTELLIGENCE AND SOFT COMPUTING
July 24-26, 2000, Banff, Canada

WORD SEGMENTATION METHOD USING INDUCTIVE LEARNING FOR CHINESE TEXT

ZHONGJIAN WANG, KENJI ARAKI and KOJI TOCHINAI

Graduate School of Engineering,

Hokkaido University

Kita 13 Nishi 8, Kita-ku,

Sapporo 060-8628 Japan

Tel, Fax: (+81-11)706-7389

{wzj, araki, tochinai}@media.eng.hokudai.ac.jp

ABSTRACT

In this paper, we describe a method that can be used to multi-language word segmentation, like Asian languages such as Japanese and Chinese. We have already used the proposed method to deal with Japanese word segmentation and got good results. In this method, we use a common character string that occurs frequently in text, and call it a common pattern. It is considered that the common pattern has a high probability as a word. It is extracted as a word candidate and registered in a dictionary. Furthermore, in our method it is not necessary to prepare a dictionary and any segmentation rules for dealing with an ambiguous segmentation beforehand. The method is also applicable to the problem of predicting unknown words and generating or augmenting existing electronic dictionaries automatically. In our method, because the grammatical rules and language knowledge are not used, it can be used to deal with other language. We have done the evaluation experiment using Chinese text of the two fields. The results of evaluation experiment show that the proposed method is also effective for Chinese word segmentation, and it is demonstrated that the proposed method is a general-purpose method for word segmentation of non-segmented languages.

Keywords: Chinese, Word segmentation, Inductive learning, Natural language processing

1 INTRODUCTION

Word segmentation of text is one of the most important technologies in computer processing of Chinese language. In the western languages, words are delimited by blanks or marks of punctuation. On the other hand, the Asian languages such as Chinese and Japanese do not mark word boundaries. However in any natural language processing application, such as machine translation and information retrieval, word segmentation of text is a very important initial stage and key technology. Because there is no blank to mark word boundaries in Chinese text, word identification is difficult, Especially, in the case that there are ambiguities in word segmentation and occurrences of unknown words which are not registered in the dictionary. A Chinese text consists of strings of Chinese characters. The same Chinese character that makes up a word may appear in different words; moreover it may be just a word [1][2]. There are mainly three kinds of methods for word segmentation of Chinese language, which are lexical rule based methods [3][4], statistical methods [5] and method of combining lexical information with statistical information [1][6][7]. The lexical rule based methods need a dictionary and rules of dealing with ambiguity segmentation. The accuracy of word segmentation greatly depends on the coverage of the underlying dictionary and the collected word segmentation rules. In addition, the identification of unknown words, the extraction and management of rules are difficult tasks. Otherwise the statistical method is by using the mutual information of characters by statistic calculation to decide the boundary of words. Generally, to construct effective model, this method needs a large amount of data and tagging corpus. The method that does not use dictionary and tagging corpus was presented [8]. However, word segmentation is still a difficult problem since ambiguous segmentation occurs.

In our method, we use a common character string that occurs frequently in text, and call it a common pattern. It is considered that the common pattern has a high probability as a word. It is extracted recursively as a word, classified into some ranks and registered in a dictionary by using inductive learning. The method of classification is based on the probability degrees. The probability degrees of the extracted common pattern are decided by the situation of extraction of the common pattern, the correct segmentation rate and the erroneous segmentation rate. Then this method divides a non-segmented text into words using the common patterns classified in order of the

higher value of the probability degrees. When there are multiple segmentations, the system gets the word candidates of possible segmentations, and picks the best segmentation from the candidates of possible segmentations by using a value of LEF calculated by the likelihood evaluation function. When the value of LEF is the same, we use the frequency of the erroneous segmentation, the frequency of the correct segmentation of word and the length of word (Maximum Length String Matching method) to decide correct segmentations, and deal with segmentation ambiguities. Furthermore, in our method, it is not necessary to prepare a dictionary and any word segmentation rules beforehand. We have applied the proposed method to Japanese text and confirmed that this method was effective for Japanese word segmentation [9]. In the proposed method, any language knowledge is not used, such as information of semantics and morphology, so that it is possible that the method is used to deal with other non-segmented languages.

The method is also applicable to predict unknown words, generate electronic dictionaries and augment existing electronic dictionaries automatically. To demonstrate the adaptability that the proposed method deals with other non-segmented languages, two fields of Chinese text were used as data for an evaluation experiment. The results show that the proposed method is effectiveness.

2 *ALGORITHM*

We show the outline of word segmentation method in Figure 1. This method consists of the following:

First, an input text is segmented by words that were acquired in the dictionary so far. This procedure is called "known word segmentation". The method of segmentation is to compare the word in the dictionary with the character string in a sentence from the beginning to the end of the sentence if both are the same, do segmentation.

Second, the remaining part of the character strings that are unsegmented by the known words, are dealt with by prediction of unknown words using inductive learning. The prediction is recursively done by extracting common patterns and high dimensional common patterns. The system does this procedure in two steps:

1. The system extracts common patterns that have the same character string of repetition in text. This step is based on the supposition that the same character string of repetition in text is highly probable to be a word.

2. The system segments the text into words using the predicted common patterns.

Third, the user judges whether the results of the word segmentation is correct or erroneous. If it is necessary, the error in results will be corrected. Then the result of segmentation and the corrected result are returned to the system.

Fourth, the system compares the corrected results with the erroneous results to renew the information of registered words. Through this procedure, the certainty that the extracted common patterns are words is confirmed and increased. Therefore the prediction ability of an unknown word is better improved. The segmentation process is done in order of high probability that the extracted common patterns are words.

2.1 *Segmentation for Known Words*

Input text and then the system segments the text into words by registered words that the system has got by using inductive learning up until that time. The method of segmentation for known words consists of two steps:

The first step, The system compares the registered common patterns in the dictionary with the character strings in the text from the beginning of the text, finds out the same character strings with the registered common patterns and segments the text into words. Then repeats this comparison process from the next character until the end of the text is reached. Here the registered common patterns are used in order of the classification of the common pattern (It is expressed at section 2.3).

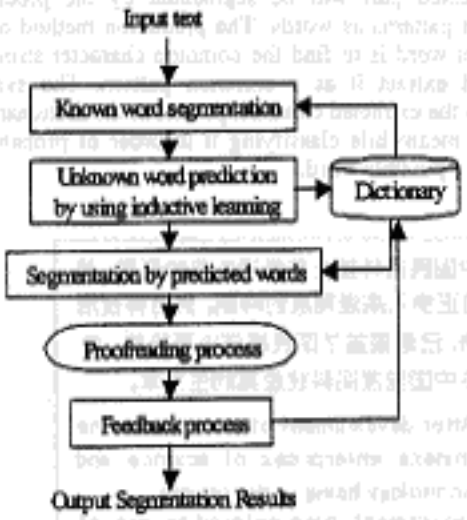

Figure 1: Outline of Word Segmentation Method Using Inductive Learning

The second step, However, for the character strings that there are several possibilities of segmentation, we first use the registered words in order of their classification in the dictionary. When there are duplicate word candidates in the same classification, we find all of the words that are possible segmentations of the text. And we decide the correct segmentation from the set of segmentation candidates by the value of likelihood evaluation function. We define the likelihood evaluation function as follows:

$$LEF(Likelihood\ Evaluation\ Function) = \alpha \cdot CS + FR - \beta \cdot ES + \gamma \cdot LE$$

Where: CS, FR, ES and LE are the frequency of the correct segmentation, the frequency of the character string appearing in the text, the frequency of the erroneous segmentation and the length of the character string respectively. α, β and γ are coefficients.

The word that has the maximum value of LEF is decided as the correct segmentation. When the LEF value of the set of possible segmentations is equal to each other, the correct segmentation candidate are decided by the word candidate that the value of ES is minimum, the value of CS is maximum, the value of FR is maximum, the value of LE is the longest or the location of segmentation is the leftmost in sentence in turn.

2.2 *Prediction for Unknown Words*

For expressing the method, an example is shown in Figure 2. Those words that are not registered in the dictionary are predicted by using inductive learning. After the sentences were segmented by known words, the unsegmented part will be segmented by the predicted common patterns as words. The prediction method of an unknown word is to find the common character string in text and extract it as a common pattern. The system registers the extracted common pattern in the dictionary as a word, meanwhile classifying it in order of probability that it is a possible word.

中国民营科技企業経過多年的発展，目前正步入高速発展的時期。民営科技活動，已経覆蓋了国民経済主要行業，成為中国発展高科技産業的生力軍。

（After development of many years, the Chinese enterprises of science and technology being under private management have entered an era of high speed development. those enterprises have become the main part of national economy and the force on development of advanced science and technology.）

Figure 2: An Example of Chinese Text

2.2.1 *Extraction of Common Pattern*

The conditions that a common pattern is extracted as a word in non-segmented text are as follows:

Condition 1: When a character string appears in non-segmented text frequently, we call it a common pattern. If the common pattern consists of more then two characters, we extract it as a word and call it S1 (segment 1). Figure 3 shows the extracted S1 from the text that is shown in Figure 2.

Extracted S1:

中国（China），国民（nation），

民営科技（private management science and technology），

科技（science and technology），

発展（develop, development）

Figure3: An Example of Extracted S1

Condition 2: When the character string appears in the text only one times but meanwhile it is included in other extracted common pattern and made up by more than two characters, we also extract it as a word. For example: 科技 (science and technology) is included in 民営科技. Therefore 科技 is extracted and belong to S1.

2.2.2 *Extraction of High Dimensional Common Pattern*

The extracted common pattern S1 at Section 2.2.1 may include other common pattern S1. At this situation, the common pattern can be extracted moreover from the extracted common pattern. This re-extraction procedure is called "Extraction of High Dimensional Common Pattern". We consider it has higher probability as a word that extracted common patterns at this procedure. The conditions of re-extraction are as follows:

Condition 1: The common patterns can be re-extracted from the extracted common patterns when it includes more than two characters. Figure 4 is an example.

民営科技（S1）includes 科技（S1）. 科技（S1）is extracted from 民営科技（S1）.

民営科技(S1) → 民営(S3) + 科技（S2）

Figure 4: An Example of Extraction of High Dimensional Common Pattern

The part of re-extraction is called S2 (segment 2); the part of remain is called S3 (segment 3). The S1 is deleted from the dictionary when it is divided into S2 and S3.

高 is in 発展高科技 and both sides of it are extracted as a word. We extract 高 (high, advanced) as a word and belong it to S2.

Figure 5: Extraction of One Character Word

Condition 2: Furthermore one character can also be extracted as a word when both sides of it are extracted as a word or both sides were segmented by known words. An example is shown in Figure 5.

2.3 *Construction of The Dictionary*

The extracted common patterns are classified to "S1", "S2", "S3" and "CW" (correct word) and registered in the dictionary. The common patterns of CW classification are common patterns that are confirmed as a word by feedback process. The construction of dictionary is like Table 1.

The FR, CS, ES, LE and CL are frequency that a common pattern appears in text, correct segmentation rate, erroneous segmentation rate, and the length of the registered common pattern and classification of the registered common patterns in the dictionary as words respectively.

Table 1: Construction of Dictionary

word	FR	CS	ES	LE	CL
中国	10	8	0	2	CW
発展	8	6	1	2	S1
国民	7	5	1	2	S1
科技	12	12	0	2	S2
高	21	14	4	1	S2
民営	5	0	2	2	S3

2.4 *Feedback Process*

After the system segments the text into words, the results are judged whether they are correct or not by the user. Then the user corrects the errors if there are errors in the results. The corrected results and erroneous results are returned to the system. The system updates the dictionary by comparing the corrected results with erroneous results.

In this process, the system updates the classification of the registered common patterns. And the system increases the priority degree of the words of correct segmentation and decreases the priority degree of words that were used in erroneous segmentations. The feedback process is described in detail as follows:

For the Results of Correct Segmentation:

When the result of segmentation is correct, the value of FR and CS of word that is used to segment are added one. If the classification of the words does not belong to CW, change it to CW.

For the Results of Erroneous Segmentation:

(1). If the dictionary does not has the correct words, the system registers the words in the dictionary, as FR of the word equals 1 and classification equals CW.

(2). If the dictionary has the correct words, the system adds one to the value of FR for a word and changes the value of CL to CW if it does not belong to CW.

(3). If the reason of erroneous segmentation is that the erroneous word was used, then the ES of erroneous word is added one.

For the Unsegmented Parts:

If the reason that the text is not segmented is that there are not words in the dictionary, the system registers the words in the dictionary as FR of the words equal 1 and classifications equal CW. Figure 6 shows an example of the result of experiment and the corrected result. The marked part of underline is an erroneous segmentation and the marked part of doubly underline is an unsegmented part.

The result of word segmentation:
/中//国民//営//科技//企業/経過/争//年//的//発展//,//目前//正//步入//高速//発展//的//時期//。//民営//科技//活動//,//已経//覆盖//了//国民//経済//主要//行業//,//成为//中国// 発展// 高//科技//産業//的//生力軍//。/
The corrected result:
/中国//民営//科技//企業//経過//争//年//的//発展//,//目前//正//步入//高速//発展//的//時期//。//民営//科技//活動//,//已経//覆盖//了//国民//経済//主要//行業//,//成为//中国//発展// 高//科技//産業// 的//生力軍//。/

Figure 6: An Example of Result of Experiment and Corrected Result

The example of the method for feedback process is shown in Figure 7.

2.5 *Management of The Dictionary*

The dictionary is generated automatically by using inductive learning and enlarges with the increase of processing text gradually. We manage the dictionary by

中//国民//営/ and 経過 are the part of erroneous segmentation and the unsegmentation part.

The process of words for correct segmentation:

科技：FR+,CS+1,the value of CL is changed to CW.

The other parts of correct segmentation is done same process

The process of words for erroneous segmentation:

中，国民 and 営：ES+1.

中国，民営：FR+1，the value of CL is change to CW.

The process of unsegmentation parts:

経過：if 経過 is not in dictionary, register it in dictionary by FR=1,CL=CW.

If 経過 is in dictionary, add one to FR of 経過 and change CL of 経過 to CW.

Figure 7: Example of Feedback Process

updating the classification of the registered words and deleting the useless words. The update of classification of word and deletion of word are decided by the value of LEF of words, as follows:

LEF ≤ −10, delete the word candidates

−10 < LEF ≤ 0, then change CL to S3

0 < LEF ≤ 10, then change CL to S1

10 < LEF, then change CL to S2

3 EVALUATION EXPERIMENTS

To confirm the effectiveness of this method for Chinese word segmentation and adaptability for data of different field, we do the experiments for data of two fields by this method.

3.1 Preliminary Experiments for Coefficients of Likelihood Evaluation Function

Before the evaluation experiment, we did the experiments to decide the optimum coefficients of the likelihood evaluation function. Selecting a series of coefficient did the experiments. We collected economics text of 200 sentences about 18,000 characters as data of preliminary experiment. The results are shown in Table 2. According to the results of experiment, we select α=1, β=50 and γ=10.

Table 2:Preliminary Experiments of Optimum Coefficients

	1	1	0	1	5	10	1	1	1	1
	1	1	1	1	1	1	50	60	70	80
	0	20	10	10	10	10	10	10	10	10
CSR[%]	84.95	86.76	86.0	86.78	84.1	82.14	87.63	87.63	87.62	87.40
ESR[%]	12.05	10.24	11.7	10.22	12.9	14.86	9.37	9.37	9.38	9.6
USR[%]	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0

3.2 Data Collection

We collect two fields of text as data of experiments from the Chinese Corpus [10]. The engineering text contains 92,085 words and the economics text contains 87,915 words. Total of the data is 180,000 words. The engineering text contains the text of electronics, communication engineering, machine engineering and nuclear industry. The economics text contains the text of economic system, economic policy and economic theory.

3.3 Experiment Procedure

At the beginning of experiment, the dictionary is empty, and input is hundred words per times. The system predicts an unknown word by a common pattern with the inductive learning. The dictionary is generated automatically along with the extraction of the common patterns as word candidates. At the feedback process, the errors in the result of word segmentation are corrected by the user. Then the corrected results and the results of segmentation containing some errors are returned to the system. The system renews the frequency of occurrence, the frequency of correct segmentation rate, the frequency of erroneous segmentation rate, and classification of the registered common patterns in the dictionary, to improve the ability of predicting an unknown word.

3.4 The Result of Experiment

To evaluate the experiment result, we define the evaluation formulas; the correct segmentation rate, the error segmentation rate and unsegmented rate are as follows:

$$\text{Correctness Segmentation Rate [\%]} = \frac{\text{Number of Correct Segmentation}}{\text{Total of Words}} \times 100$$

$$\text{Eorror Segmentation Rate [\%]} = \frac{\text{Number of Eorror Segmentation}}{\text{Total of Words}} \times 100$$

$$\text{Unsegmentation Rate[\%]} = \frac{\text{Number of Unsegmentation}}{\text{Total of Words}} \times 100$$

The results of experiment are shown in Table 3. The average correct segmentation rate is 88.6%, the average erroneous segmentation rate is 5.3% and the average unsegmented rate is 6.2%. This result are got under a state of the dictionary is empty at beginning. In the Figure 8, 9 and 10, the change of the correctness segmentation rate, the unsegmented rate and the error segmentation rate are demonstrated respectively. In these Figures, the data of experiment is made up of two parts, engineering text and economics text.

Table 3: Result of Experiment

CSR[%]	ESR[%]	USR[%]
88.6	5.3	6.2

4 DISCUSSION

When number of processed words is about 5000 in Figure 8, the correctness segmentation rate is near to 100%. And after 5000 words, the correctness segmentation rate is going down because of ambiguities, however the correctness segmentation rate is going up along with increasing of processed words. CL of word candidates is updated and adjusted uninterruptedly. When the text is changed to different domain, the correctness segmentation rate is fall down temporary. But goes on increasing quickly. Sometimes the correctness segmentation rate is a little lower because the domain of text is a little difference.

We did the experiments with text of two fields, and got the acceptable correctness segmentation rate. The experiments were done at the beginning of empty of the dictionary. The experiment results show the predictive ability of an unknown word by using the inductive learning. We understand that the proposed method is effective for prediction of the unknown words. We use only the information of character strings of text in this method. Because any language knowledge is not used, the proposed method can be used on other non-segmentation language.

At feedback process, a user judge the result whether it is correct or not, and corrects the errors in result. As the result of the deviation of judgment of a user for segmentation standard, the correct segmentation rate was perhaps affected some degrees. If the amount of data of experiment is increased, the unsegmented rate will be going down and the correctness segmentation rate will be improved.

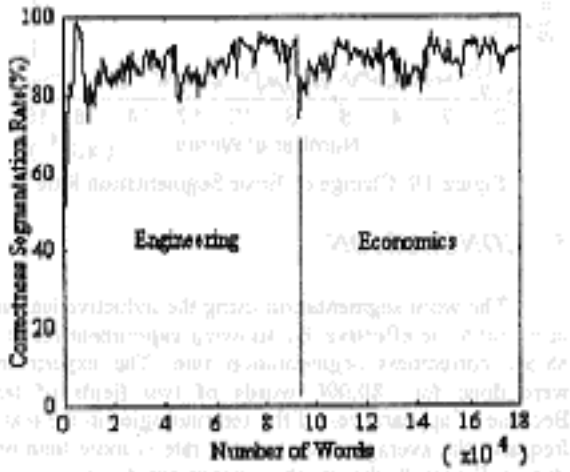

Figure 8: Change of Correctness Segmentation Rate

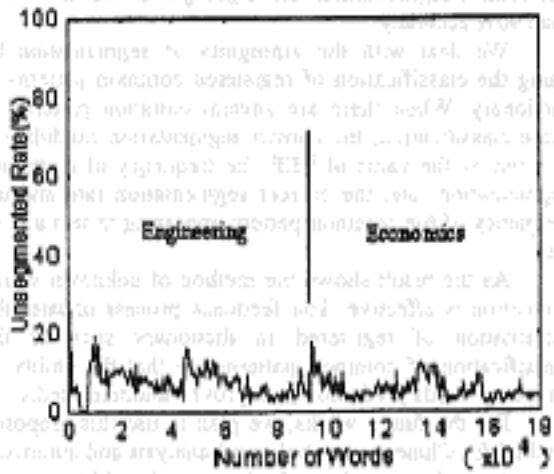

Figure 9: Change of Unsegmented Rate

In the Figure 8, when the data is changed from engineering text to economics text, the correctness segmentation rate has a little decrease, but goes on increasing quickly. The correctness segmentation rate of the result has no evident variation, so that we can consider that the proposed method has also adaptability for different fields.

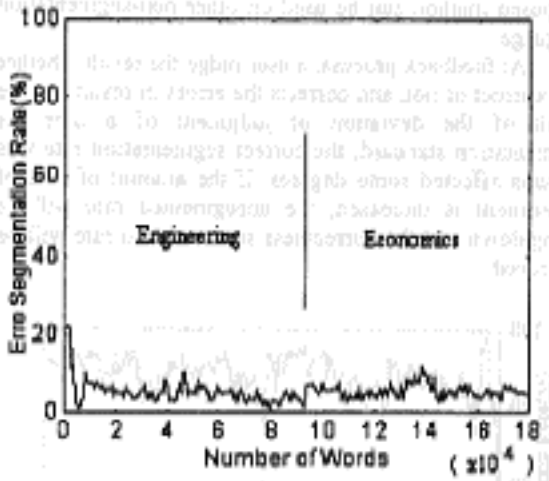

Figure 10: Change of Error Segmentation Rate

5 CONCLUSION

The word segmentation using the inductive learning turns out to be effective, by showing experiment result of 88.6% correctness segmentation rate. The experiments were done for 180,000 words of two fields of text. Because of appearances of the terminologies in the text is frequent, the average unsegmented rate is more than 6%. But as Figure 9 shows, the unsegmented rate is going down with increasing of the processed text. The correctness segmentation rate is going to be stable at more than 90% accuracy.

We deal with the ambiguity of segmentation by using the classification of registered common patterns in dictionary. When there are several common patterns of same classification, the correct segmentation candidate is decided by the value of LEF, the frequency of erroneous segmentation rate, the correct segmentation rate and the frequency of the common pattern appearing in text and so on.

As the result shows the method of unknown words prediction is effective. The feedback process updates the information of registered in dictionary such as the classification of common patterns, so that the ability of unknown words prediction is improved uninterruptedly.

For the future works, we plan to use this proposed method for Chinese morphological analysis and automatic generation of terminology dictionaries. In addition, more detailed evaluation is necessary.

ACKNOWLEDGEMENTS

We acknowledge the use of Sinica Corpus in evaluation experiment.

REFERENCES

[1] Richard Sproat, Chilin Shih, William Gale, and Nancy Chang, A stochastic finite-state word-segmentation algorithm for Chinese, *Association for Computational Linguistics*, 22(3), 1996, 377-404.

[2] Zhiyuan Chen, Improve of Chinese language software by morpho-syntactic and lexicographic methods, *PACLING'99, WATERLOO, CANADA*, 1999, pages 295-301.

[3] GuPing, WuTao, and Mao Yuhang, The adjacent matching algorithm of Chinese automatic word segmentation and its implementation in QHFY Chinese-English system, *Computational Linguistics Research and Application*, pages, 1993, 132-138 (in Chinese).

[4] Sheng Yuan Wu, A new Chinese phrase segmentation method, *Computer Research and Development*, 33(4), 1996, 306-311(in Chinese).

[5] Chaojan Chen, Ming hong Bai, and Kehjian Chen, Category guessing for Chinese unknown words, *Proceedings of NLPR'97, Phuket, Thailand, December 2-4*, 1997, pages 35-40.

[6] Wanying Jin, Chinese segmentation disambiguation, *The 15th International Conference on Computational Linguistics*, 1994, pages 1245-1249.

[7] Ron Song, Hong Zhu, Weigui Pan, and ZhenhaiYin, Automatic recognition of person names based on corpus and rule-base, *Computational Linguistics Research and Application*, 1993, pages 150-154 (in Chinese).

[8] Maosong Sun, Dayang Shen, and Benjamin K Tsou, Chinese word segmentation without using lexicon and hand-crafted training data, *17th International Conference on Computational Linguistics*, 1998, pages 1265-1271.

[9] Kenji Araki, Yoshio Momouchi, and Koji Tochinai, Evaluation for Adaptability of Kanji-Kana Translation of Non-Segmented Japanese Kana Sentences Using Inductive Learning, *In Proceedings of the Second Conference of the Pacific Association for Computational Linguistics*, April, 1995, pages 1-7.

[10] Sinica Corpus, http://rocling.iis.sinica.edu.tw// ROCLING/index_c.htm