Generality for Multi-language of Word Segmentation Method Using Inductive Learning

Zhongjian Wang, Kenji Araki and Koji Tochinai
Graduate School of Engineering, Hokkaido University
Kita 13 Nishi 8, Kita-ku, Sapporo 060-8628, Japan
Email: {wzj,araki,tochinai}@media.eng.hokudai.ac.jp

Abstract

We have proposed a method of word segmentation for non-segmented language using inductive learning. We use only surface information of a character string, so that the method has an advantage that is entirely not dependent on any specific language. We have confirmed its effectiveness for Japanese and Chinese word segmentation respectively. In this paper, we will discuss the generality of our proposed method for multi-language. We used a large amount of experimental data from Japanese corpus EDR and Chinese corpus Sinica to carry out the evaluation experiments. For these two kinds of language that they are quite different on grammar, structure and morphology, we have used the same algorithm to carry out the evaluation experiments. The experimental results show our proposed method is effective for Japanese and Chinese word segmentation, and it is possible to be used to multi-language.

Keywords: multi-language, inductive learning, word segmentation, generality.

1 Introduction

In computer processing of non-segmented language, like Chinese, Japanese and Thai, word segmentation for text is one of the most important technologies. Unlike English and other western languages, the non-segmented languages do not mark word boundaries. A sentence is composed of continuous character strings without space between words. However in any NLP application, word segmentation of text is a very important initial stage. There are mainly three kinds of methods for word segmentation of non-segmented language, which are lexical rule based methods [1][2][3], statistical methods [4][5][6][7] and method of combining lexical information with statistical information [8]. The lexical rule based methods and the method of combining lexical information with statistical information need a dictionary and rules to deal with ambiguous segmentation. The accuracy of word segmentation greatly depends on the coverage of the underlying dictionary and the collected rules of segmentation. In addition, the identification of unknown words, the extraction and management of rules are difficult tasks. Otherwise the statistical method is using the mutual information of characters by statistic calculation, to decide the boundary of words. Generally to construct effective model avoiding the problem of sparseness, this method needs a large amount of data.

In [3], authers make use of the common characterastic of different language to propose a language independent morphological analysis system. In this method, a set of morphological element for fixed language is prepeard by a user and a common method is used in searching the optimum morphological candidate for different language. In [6], A statistical method that uses raw text data as training data by calculating mutual information and t-score between Chinese characters to decide boundaries of words is proposed. The [7] proposed a statistical method utilizing unsegmented training data for segmentation of Kanji sequences in Japanese text, by calculating n-gram to decide boundaries of words. However evaluation experiments are not done for Chinese language.

With the development of internet and popularization of computers, a large amount of text information in different languages on internet are increasing explosively, so that it is necessary to develop a common method to deal with multi-language.

In our method, we extract recursively a common part and a different part of a character string that occur frequently in text as word candidates. Those extracted common parts and different parts are classified into some ranks according to extracting condition, and registered in a dictionary. We consider that the common part and the different part have a high probability as a word. The proposed method segments a non-segmented sentence into words using the ranks of common parts and different parts in order of the higher value of the certainty degrees. When there are multiple possible segmentation, the system gets all candidates of possible segmentation, and picks a correct segmentation from the candidates by using a value of LEF (the likelihood evaluation function, Section **2.1**). When the LEF value of candidates is the same, we use the frequency of the erroneous segmentation, the frequency of the correct segmentation

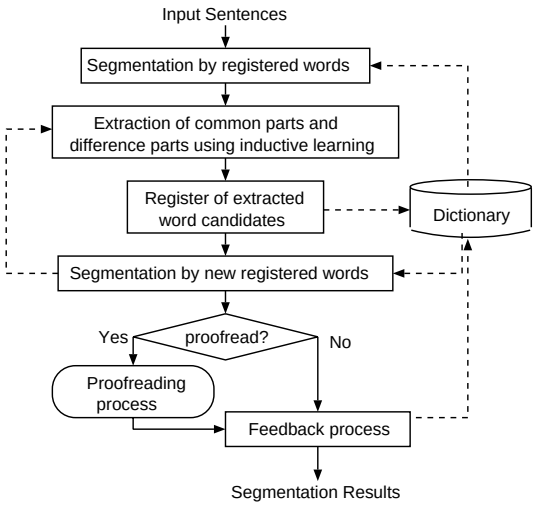

Figure 1: Overview of the word segmentation system

and the length of word to decide a correct segmentation, and deal with segmentation ambiguities. In our method, it is not necessary to prepare a dictionary and any word segmentation rules beforehand. Because only surface information of a character string is used, it is possible that our method is used to deal with general non-segmented language.

Japanese language is quite different from Chinese language on structure, grammar, morphology and syntactic characteristics. Japanese sentences are writen by two kinds of character, about several thousands Kanji and fifty Kana. On the other hand, Chinese sentences are written only by Chinese character, there are about fifty thousands Chinese characters. The structure of sentence is also different, for example: structure of a Chinese sentence is SVO (subject+verb+object) but structure of a Japanese sentence is SOV. A word is comprised at least one Chinese character. A Chinese character may be a part of a word, and may be appear in other word, and may be just a word. So that ambiguity segmentation occurs very easily. We have processed these two kinds of language only as a general character symbol, and have got satisfy results. By the results, the proposed method is confirmed that it is possible to be used to general non-segmented language.

2 Overview of the Proposed Method

Fig.1 shows the outline of the word segmentation method. This method consists of the following:

(1). An input sentence is segmented by words that were acquired in the dictionary so far. This procedure is called "segmentation by known words". 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, to find all of words that are usable to segment and make a list of segmentation candidates. The words that have high value of LEF(Section **2.1**) are first used for segmentation at first.

(2). 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". A character string of apearing repetitionary in text is extracted and called a common part. Between those extracted common parts, may still have a common character string. We extract the common character string, and call them common parts and call the remain part different parts respectively. The prediction of an unknown word is recursively done by extracting common parts and different parts of a character string. Furthermore the extraction process proceeds on two stages: extraction of common part, extraction of high dimensional common part and different part. A high dimensional common part is a common part that is extracted by more than two times. We consider a high dimensional common part has high probability as a word.

The system extracts common parts and different parts of a common character string that appears in a sentence repeatedly. This process is based on the supposition that a common character string of repetition in text has highly probability to be a word. After extracting common parts and different parts, they are registered in dictionary. Then the system segments the text into words using the extracted common parts and different parts.

(3). The user judges whether the results of the word segmentation is correct or not. If there are errors in the segmentation result, the user will correct errors. Then the result of segmentation and the corrected result are returned to the system.

(4). The system compares the corrected results with the segmentation results to renew the information of registered words in the dictionary. Through this procedure, the certainty that the extracted common parts and different parts are as words is confirmed and increased. The segmentation process is done in order of high probability that the extracted common parts and different parts are as words.

Here, the extracted common part and different part are generally named WS (Word Segment). And the WS those are used in correct segmentation are generally named CW (Correct Word).

2.1 Segmentation by Known Words

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

(1). In first step, the system compares the registered CW and WS in the dictionary with the character

string in the sentence from the beginning of the sentence, finds out the same character strings with the registered words, and repeats this comparison process until the end of the sentence is reached. A list of segmentation candidate is established. Then the system segments the sentence into words. Here a correct candidate are used to segment in order of the classification of them (Section **2.3**).

(2). In 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 is more than one word candidate with the same classification, we decide the correct segmentation from the list of segmentation candidates by the value of likelihood evaluation function. We define the likelihood evaluation function(LEF) as follows:

$$LEF \quad = \quad FR + \alpha CS - \beta ES + \gamma LE \qquad (1)$$

Where: FR, CS, ES and LE are the frequency of a common part or a different part appearing in the text, the frequency of the correct segmentation, the frequency of the erroneous segmentation and the length of a common part or a different part respectively. α, β and γ are coefficients.

The word that has the maximum value of LEF is decided as the correct segmentation candidate.

(3). When the LEF value of the set of possible segmentations is equal to each other, the correct segmentation candidate is 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 a sentence in turn.

2.2 Prediction for Unknown Words

For expressing the method, an example is shown in Figure 2. In this example, every letter represents a Chinese character or a Kanji, so we use this example to express a general sentence of non-segmented language for explanation our method. Those words that are not registered in the dictionary are predicted by using the inductive learning. After the sentences were segmented by CW and WS, which have been registered in the dictionary, the unsegmented part of character string will be segmented using the extracted common parts and different parts. The prediction method of an unknown word is to find the common character string in text and extract it. Sometimes the extraction procedure proceeds on two stages: the extraction of common parts, the re-extraction of common parts and the extraction of different parts. So that the common character string is extracted as a common part and a remaining part is as a different part. The system registers the extracted common part and different part in the dictionary as a word candidate, meanwhile

$$\underline{\alpha\beta\chi\delta}\kappa\varphi\underline{\epsilon\phi\gamma\Theta\pi\mu\tau\gamma\beta}\pi\underline{\alpha\beta\chi\delta}$$
$$\underline{\Theta\pi\mu\tau\gamma\beta}\eta\Psi\underline{\epsilon\phi\gamma\tau\gamma\beta}\alpha\zeta\theta.$$

Figure 2: An example of non-segmented sentence.

classifying it in order of probability that it is possible as a word.

2.2.1 Extraction of Common Part

The extraction of a common part in non-segmented text is two steps:

(1). When a character string appears in non-segmen ted text frequently, we call it a common character string. If the common character string consists of more than two characters, we extract it as a word candidate and call it common part and represent it by S1 (Segment one). The extracted S1 from the sentence that is shown in Fig. 2: "$\alpha\beta\chi\delta$","$\epsilon\phi\gamma$" and "$\Theta\pi\mu\tau\gamma\beta$".

(2). When the character string appears in the sentence only one times but meanwhile it is included in other extracted common part and made up by more than two characters, we also extract it as a word candidate. For example in Fig. 2: "$\tau\gamma\beta$" is included in "$\Theta\pi\mu\tau\gamma\beta$". Therefore "$\tau\gamma\beta$" is extracted and belong to S1.

2.2.2 Extraction of a High Dimensional Common Part and a different Part

The extracted common part S1 at **2.2.1** may still include other common character string. At this situation, the common character string can be re-extracted moreover from the extracted common part S1. We consider it has a higher probability as a word that re-extracted common parts at this procedure. The conditions of re-extraction are as follows:

(1). The common parts can be re-extracted from the extracted common part S1 when it includes a common character string that is more than two characters. For example, "$\Theta\pi\mu\tau\gamma\beta$" contains "$\tau\gamma\beta$"; "$\tau\gamma\beta$" can be extracted from "$\Theta\pi\mu\tau\gamma\beta$", i.e. : "$\Theta\pi\mu\tau\gamma\beta$(S1)" = "$\Theta\pi\mu$(S2)" + "$\tau\gamma\beta$(S3)".

The part of re-extraction is called "High Dimensional Common Part" and is represented by S2 (Segment two); the part of remain is called "different part" and is represented by S3 (Segment three). The S1 is deleted from the dictionary when it is divided into S2 and S3.

(2). Furthermore one character can also be extracted as a word candidate when both sides of it are extracted as a word candidate or both sides were segmented by known words. Like "π" in "$\Theta\pi\mu\tau\gamma\beta\pi\alpha\beta\chi\delta$"

Table 1: Construction of the dictionary.

Word	FR	CS	ES	LE	CL
$\alpha\beta\chi\delta$	10	8	0	4	CW
$\tau\gamma\beta$	12	12	0	4	S2
π	21	14	4	2	S2
$\epsilon\phi\gamma$	8	6	1	4	S1
$\Theta\pi\mu$	7	5	1	4	S3

is surrounded by "$\Theta\pi\mu\tau\gamma\beta$" and "$\alpha\beta\chi\delta$", and "$\pi$" is extracted as a word candidate belonging to S2.

2.3 Construction of the Dictionary

The extracted common parts and different parts are represented by WS (Word Segment), classified to "S1", "S2", and "S3". In the dictionary, "CW" (Correct Word) is WS that are confirmed as a word by proofreading process in segmentation processing. The construction of a dictionary is like Table 1.

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

2.4 Feedback Process

After the system segments the sentence 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 information in the dictionary by comparing the corrected results with the erroneous results. In this process, the system updates the classification of the registered CW and WS. And the system increases the priority degree of the words that were used in correct segmentation and decreases the priority degree of words that were used in erroneous segmentations. The feedback process is described in detail as follows:

(1) 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.

(2) For the Results of Erroneous Segmentation:

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

(3) For the Unsegmented Parts:

- The system registers the words in the dictionary as FR of the words equal 1 and classifications equal CW.

3 Evaluation Experiments

We do evaluation experiments using Japanese text and Chinese text from EDR corpus and Sinica corpus respectively. We use the Chinese text of three specialized fields to evaluate adaptivity of the proposed method for different field text. We also use two kinds of language: Japanese text and Chinese text, to evaluate generality of method for adapting different languages.

3.1 Preliminary Experiments for Coefficients of Likelihood Evaluation Function

Before the evaluation experiment, we did the preliminary experiments to decide the optimum coefficients of the likelihood evaluation function use greedy method. We let α change and give β and γ any a value of constant, decide value of α when the result of experiment is the best. Then we change β; fix α and γ as a constant. We repeat this procedure until the result of experiment is no big change. We collected economics text of 250 sentences about 7,000 words as data of preliminary experiment[9]. According to the results of experiment, we decide optimum coefficients α=1, β=50 and γ=70.

3.2 Experimental Data

We collect three fields of text from the Sinica Corpus[1] : the architecture text contains 145,727 words, the economics text contains 113,000 words and the electronic engineering text contains 116,110 words. Total of the data is 374,837 words. The architecture text consists of architecture report, architecture news and architecture aesthetics. The electronic engineering text consists of the text of electronics, communication engineering, machine engineering and nuclear industry. The economics text consists of the text of economic system, economic policy and economic theory.

We select Japanese text about 389,230 words from EDR[10]. the Japanese text is not classified by fields.

[1] $http://www.sinica.edu.tw/ftms-bin/kiwi.sh$

Table 2: Experimental result of Japanese text and Chinese text

Language	Chinese text(total words : 374,837)				Japanese text
Field	Architecture	Economics	Electronic-engineering	Average	
Number of words	145,727	113,000	116,110	374,837	389,230
CSR[%]	89.3	91.2	91.9	90.6	87.5
ESR[%]	10.1	8.4	8.1	9.0	12.3
USR[%]	0.6	0.4	0.0	0.4	0.2

3.3 Experiment Procedure

At the beginning of experiment, in order to confirm that the proposed method is a language independent method, we let the initial dictionary empty. A dictionary is generated along with extraction of common parts and different parts in text. An input is a paragraph about hundred words one times. The system predicts an unknown word by extracting a common part and different part using the inductive learning. The dictionary is generated automatically along with the extraction of common parts and different parts as word candidates.

At the feedback process, a user corrects the errors in the result of word segmentation. 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, the frequency of erroneous segmentation, and classification of registered common parts, different parts in the dictionary, to improve the ability of predicting an unknown word.

3.4 Results of Experiment

In our method, the correct segmentation number is the number of correct segmentation that is judged by a user. The unsegmentation number is the number when all unsegmented strings are segmented correctly. The erroneous segmentation number is the number that all word number in input text reduces the correct segmentation number and reduces the unsegmentation number. To evaluate the experiment result, we define the evaluation formulas; CSR (Correct Segmentation Rate), ESR (Erroneous Segmentation Rate) and USR (Unsegmented Rate) are as follows:

$$CSR[\%] = \frac{Correct\ segmentation\ number}{Total\ number\ of\ words} \times 100 \quad (2)$$

$$ESR[\%] = \frac{Erroneous\ segmentation\ number}{Total\ number\ of\ words} \times 100 \quad (3)$$

$$USR[\%] = \frac{Unsegmentation\ number}{Total\ number\ of\ words} \times 100 \quad (4)$$

The results of experiment are shown in Table 2. For Chinese text, the average correct segmentation rate is 90.6%, the average erroneous segmentation rate is 9.0% and the average unsegmented rate is 0.4%. For Japanese text, the average correct segmentation rate is 87.5%, the average erroneous segmentation rate is 12.3% and the average unsegmented rate is 0.2%. This result are got under a state of the initial dictionary is empty.

Fig.3 and Fig.4 show the change of the correct segmentation rate, the unsegmented rate and the erroneous segmentation rate of Chinese text and Japanese text respectively.

4 Discussion

4.1 Adaptivity for Different Fields

Fig.3 shows the experimental results of three fields. When the text is changed to different domain, because appearance of some new words of different domains, the correct segmentation rate is fall down temporary. However with increasing of processed amount of word, the correct segmentation rate goes on increasing quickly.

We may consider that the proposed method has adaptability for different fields. Sometimes the correct segmentation rate is a little lower because the domain of text is a little difference, for example: the architecture text consists of architecture report, architecture news, architecture aesthetics and so on.

4.2 Generality of Adapting to Different Languages

By comparing the Chinese experimental result Fig.3 with the Japanese experimental result Fig.4, we can say the proposed method is effective for both Chinese word segmentation and Japanese word segmentation. For these two kinds of quite different languages, we got the experimental results of 90.6%(Chinese text) and 87.5%(Japanese text) correct segmentation rate by using the same method.

The correct segmentation rate of Japanese text is a little lower than Chinese text, which is because text of EDR corpus is not classified by a field. Since sentences

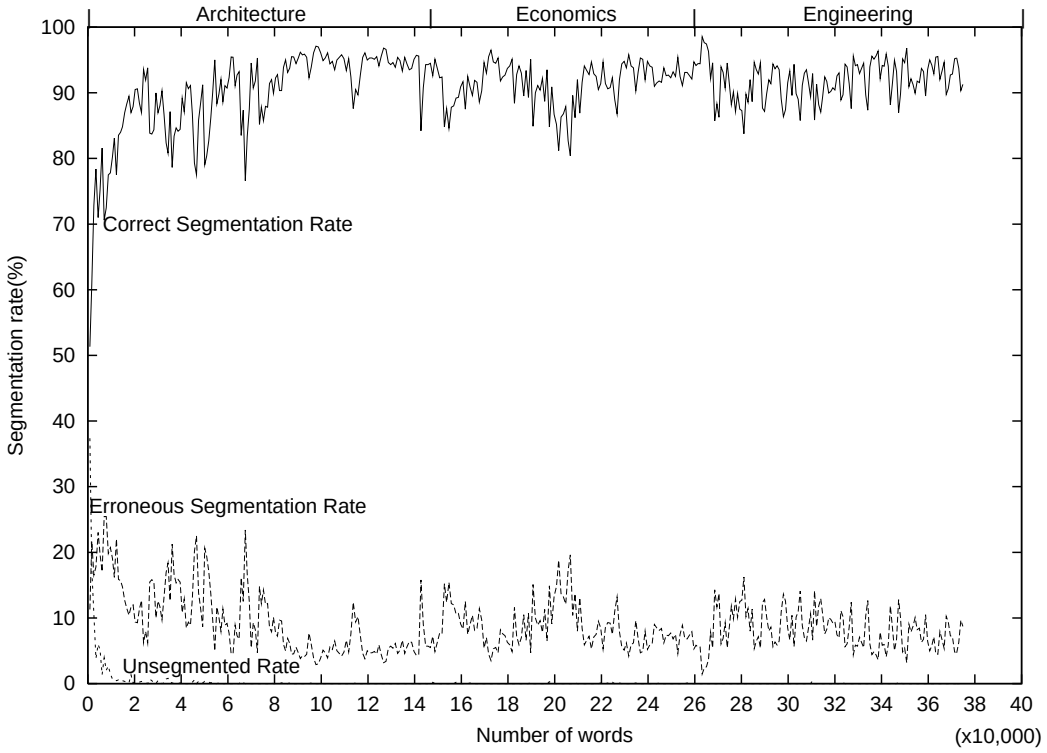

Figure 3: The change in segmentation rate of Chinese text

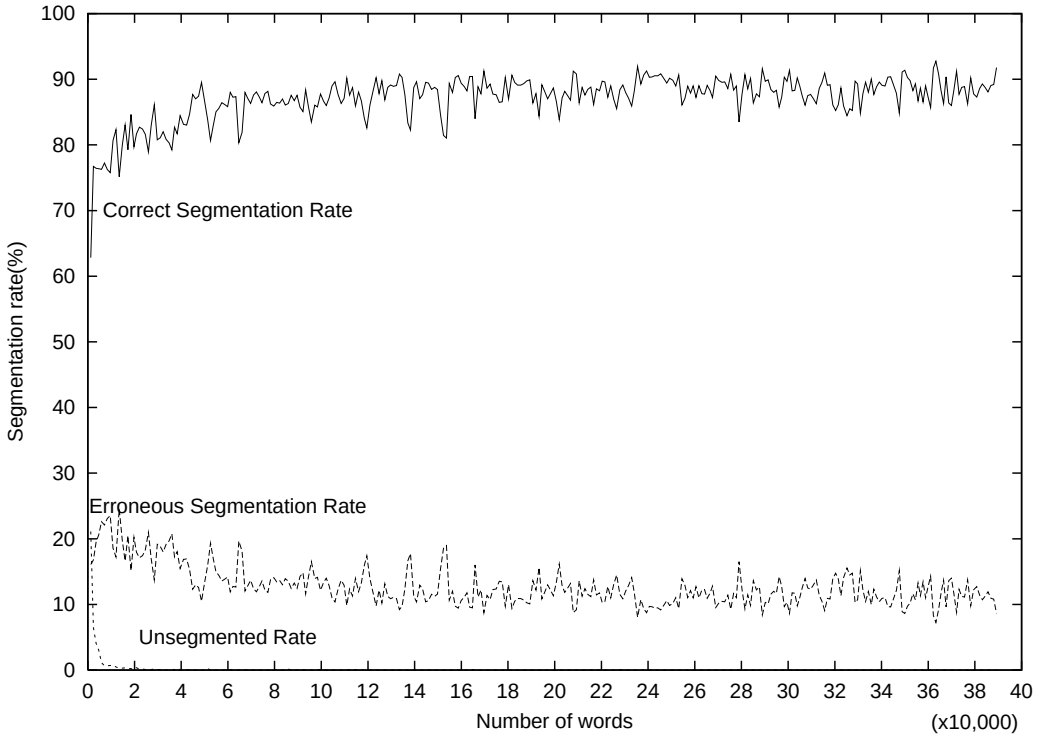

Figure 4: The change in segmentation rate of Japanese text

Table 3: Ratio of error segmentation

Range of words		0-1,000	360,000-361,000
Ratio	ESR of unregister	15.0%	0.8 %
	ESR of ambiguity	1.5%	6.0 %
	ESR of corpus jolt	0.5%	0.4 %
Total	Average ESR	17.0%	7.2 %

of a different field are mixed together, the correct segmentation rate in Fig.4 is relatively smooth. However Fig.4 shows the rising tendency of the correct segmentation rate, the average correct segmentation rate of all segmentation result is 87.5%.

We consider the proposed method is an effective method for dealing with other non-segmented language and it is a general-purpose method.

4.3 Evaluation of Ability for Predicting Unknown Words

After system processes about 45,000 words, there are almost not unknown words. We select 50,000 words to discuss the predicting ability of proposed method for unknown words.

$$Precision[\%] = \frac{CWN}{TWN} \times 100 \quad (5)$$

$$Recall[\%] = \frac{CWN}{TUN} \times 100 \quad (6)$$

Where, CWN is the number of words that are predicted correctly. TWN is the total number of words that are predicted. TUN is the total number of unknown words.

The precision and recall are shown in Fig.5. The average precision is 26.0%. The average recall is 31.0%. With incresing of registered words in the dictionary, prediction effect for unknown words is becoming well, after 40,000 words are processed the precision and the recall are 85.0%, 40.0% respectively.

4.4 Analysis of Erroneous Segmentation

We select 1,000 words from beginning of the text and between 360,000 to 361,000 words respectively, to analysis the reason of an erroneous segmentation. The results are shown in Table 3. At the beginning, ESR that is because of unregistered words is 15.0%, but after 361,000 words are processed, ESR that is because of unregistered words is 0.8%. However ESR that is caused by ambiguity goes on increasing from 1.5% to 6.0%. ESR caused by ambiguity is increasing with increasing of registered word in the dictionary. Ambiguous segmentation is still a difficult problem, so that it is necessary to improve the ability to deal with ambiguity.

Table 4 shows proofreading times in experiment of Chinese text. Fig. 6 shows the change in proofreading times of errors in segmentation result of Chinese and Japanese text. The proofreading times includes correct times for segmentation errors, segment times for unsegmentation part. According to Fig. 6 shows, the proofreading times is reducing sharply with increasing of processed text.

4.5 Comparing Experiment with Other Approach

We have done the comparing experiment with other approach to illustrate the effectiveness of the proposed method. In [6], the method of Chinese word segmentation without using lexicon and hand-crafted training data was proposed. The method decides boundaries of words by calculating mutual information and the difference of t-score between characters. They did evaluation experiment by using 20Mbyte of text as training data and 100 sentences as test data. We have completed this system for comparing experiment, and use architecture text of 512Kbyte (145,727 words) as training data. We take out 100 sentences from training data as experimental data. We used the segmentation result of Sinica corpus as standard of correct answer. As Table 5 shows, we evaluated the experimental result by evaluation method of [6] and our evaluation method [2] respectively. According to the experimental results, the proposed method is effective, and can get good results with a little of data.

5 Conclusion

We did the experiments with Chinese text of three fields and Japanese text, and got the acceptable correct segmentation rate. The experiments were carried out 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. The experiment results of three fields shown the proposed method can adapt to different fields text, and the experiment result of two kinds of language shown the proposed method is independent of language. Furthermore we use only the information of character strings of text in this method. According to the experiment results and any language knowledge is not used in proposed method, so it may be used on other non-segmentation language.

The word segmentation using the inductive learning turns out to be effective, by showing experiment results. The experiments were carried out for Chinese text of three fields and Japanese text. The correct

Table 4: Proofreading times.

Range of Word	0∼1,000	10,000∼11,000	20,000∼21,000	30,000∼31,000
Proofreading Times	98	132	84	96
Range of Word	40,000∼41,000	50,000∼51,000	60,000∼61,000	70,000∼71,000
Proofreading Times	89	68	44	45
Range of Word	80,000∼81,000	90,000∼91,000	100,000∼101,000	110,000∼111,000
Proofreading Times	87	45	36	44

Table 5: Results of comparison experiment.

Experiment results	by method of [6]	by our method	by our method
Evaluation method	by method of [6]	by method of [6]	by our method
CSR[%]	89.6	97.8	94.6

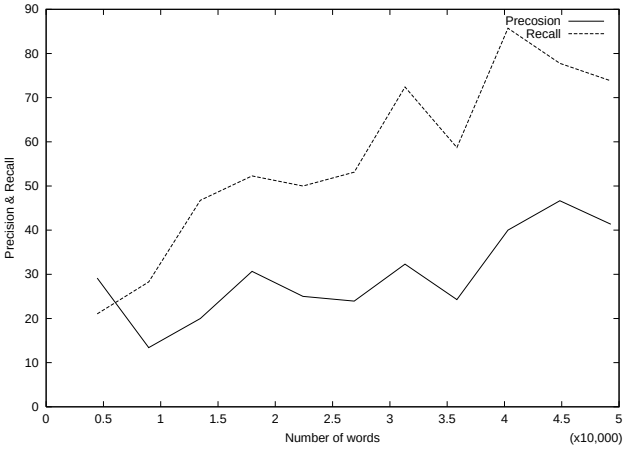

Figure 5: The ability to predict unknown words

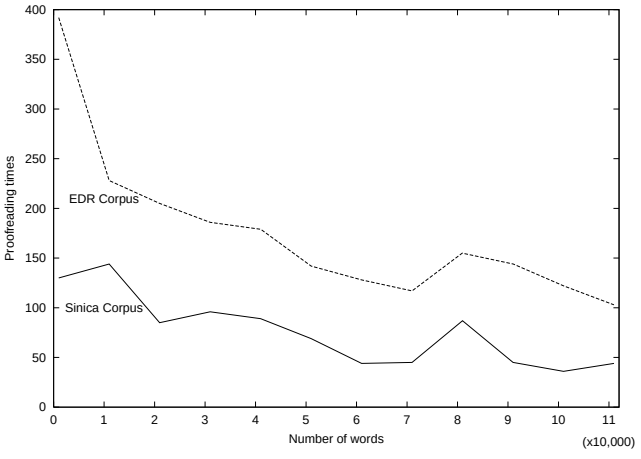

Figure 6: The change in proofreading times

segmentation rate in Fig. 3, Fig. 4 are going to be stable at more than 95% and 90% respectively.

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

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 parts and different parts, 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.

References

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

[2] Chen, K.J and S.H. Lin, "Word Identification for Mandarin Chinese Sentences," Proceedings of Coling 92, Nantes, France, August 1992, pp.101-107.

[3] TATUO TAMASITA and YUJI MATSUMOTO, "Framework for Language Independent Morphological Analysis," Journal of Natural Language Processing, Vol.7, No.3, July 2000, pp. 39-56(in Japanese).

[4] Sproat, R. and Shih, C., "A statistical method for finding word boundaries in Chinese text," Computer Processing of Chinese and Orien-tal Languages, Vol.4, No.4, 1990, pp. 336-351.

[5] Hiroki ODA, and Kenji Kita, "A Japanese Word Segmentation Using a PPM*-based Language Model," Information processing, Vol.41, No.3, Mar. 2000, pp. 689-700(in Japanese).

[6] 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, pp. 1265-1271.

[7] Rit Kubota and Lillian Lee,b "Mostly- Unsupervised Statistical Segmentation of Japanese: Applications to Kanji," ANLP-NAACL 2000, pp.214-248.

[8] Sproat, R., Shih, C., W.Gale and N.Chang, "A stochastic finite-state word-segmentation algorithm for Chinese," Computational Linguistics, Vol.22, No.3, pp. 377-404,1996.

[9] Zhongjian Wang, Kenji Araki and Koji Tochinai, "Word Segmentation Method Using Inductive Learning for Chinese Text," Proceedings of the IASTED International Conference, Artificial Intelegence and Soft Computing, July 2000, pp.452-458.

[10] Japanese Electronic Dictionary Research Institute, Ltd. (1993). EDR Electronic Dictionary Specification Guide (in Japanese).