A Word Segmentation Method With Dynamic Adapting To Text Using Inductive Learning

Zhongjian Wang
Hokuto System Co.,LTD
Oyachi Higashi1-3-23, Atubetuku, Sapporo, 004-0041 Japan
wang@hscnet.co.jp

Kenji Araki
Graduate School of Engineering, Hokkaido University
N13-W8, Kita-ku, Sapporo, 060-8628 Japan
araki@media.eng.hokudai.ac.jp

Koji Tochinai
Graduate School of Business Administration, Hokkai-Gakuen University,
Asahimachi 4-1-40, Toyohira-ku, Sapporo, 062-8605 Japan

Abstract

We have proposed a method of word segmentation for non-segmented language using Inductive Learning. This method uses only surface information of a text, so that it has an advantage that is entirely not dependent on any specific language. In this method, we consider that a character string of appearing frequently in a text has a high possibility as a word. The method predicts unknown words by recursively extracting common character strings. With the proposed method, the segmentation results can adapt to different users and fields. To evaluate effectivety for Chinese word segmentation and adaptability for different fields, we have done the evaluation experiment with Chinese text of the two fields.

1 Instruction

In NLP applications, word segmentation of non-segmented language is a very necessary initial stage(Sun et al., 1998). In the other hands, with the development of the Internet and popularization of computers, a large amount of text information in different languages on the Internet are increasing explosively, so it is necessary to develop a common method to deal with multi-language(Yamasita and Matsumoto, 2000). Furthermore, the standard of word segmentation is dependent on a user and destination of use(Sproat et al., 1996), so that it is necessary that word segmentation can adapt users, can deal with multi languages.

In our method, we extract recursively a common character string that occur frequently in text and call it a common part. When some common parts contain still same character strings, furthermore we extract the same character string as high dimensional common parts and the remain parts is called different parts. The high dimensional common parts maybe have higher possibility as words because it is extracted by multi steps. Those extracted common parts and different parts are called WS(Word Segment), and classified into some ranks according to extracting condition. The proposed method segments a non-segmented sentence into words using the ranks of WS in order of the higher value of the certainty degrees as words. When there are multiple segmentation candidates, the system gets a list of segmentable candidates, and picks a correct segmentation candidate from the list by using a value of LEF (Likelihood Evaluation Function, Section **2.1**) and so on. In addition, it is not necessary to prepare a dictionary and any word segmentation rules beforehand. A dictionary of adapting to the user or the field is generated with increasing of processed text. Because only surface information of a text is used, it is possible the method is used to deal with general non-segmented language. Here Inductive Learning is the procedure to extract recursively WS by multi steps(Araki et al., 1995).

2 Algrithm

Fig. 1 shows the outline of the proposed method.

(1) Input sentences are segmented by word candidates that were acquired in the dictionary so far.

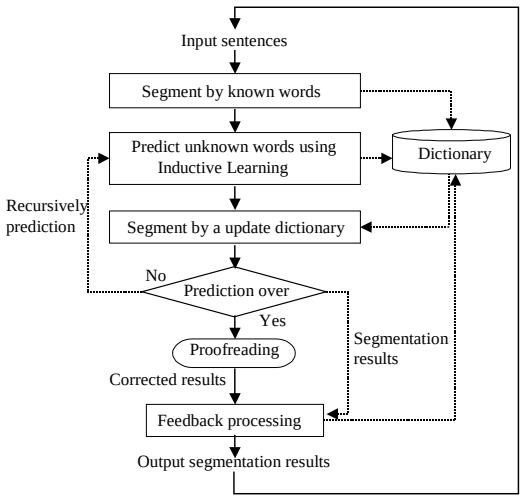

Figure 1: Outline

(2) For the remaining part of the character strings that are unsegmented by the known words, the system predicts unknown words by extracting WS using Inductive Learning.

The system extracts WS as word candidates. This process is based on the supposition that a common character string of appearing repeatedly in text has high probability as a word.

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

(4) The system compares the proofread results with the segmentation result to update the information in the dictionary. Through this procedure, the certainty of WS as a word is confirmed and increased.

Here, the WS those are used in correct segmentation are called CW (Correct Word).

2.1 Segmentation by Known Words

Input sentence and then the system segments it into words by registered CW and WS that the system has got by using Inductive Learning until that time.

(1) In the first step, the system compares the registered CW or WS in the dictionary with the character string in the input sentence from the beginning of the sentence, and finds out the same character strings with the registered words. The system 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.

(2) In the second step, however, for the character strings of multiple segmentations, we use the registered candidates in order of their ranks in the dictionary(Section **2.3**). When there are more than one word candidate with the same rank, we decide the correct segmentation from the list of segmentation candidates by the value of LEF. We define LEF as follows:

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

Where: FR, CS, ES and LE are the frequency of CW or WS appearing in the text, the frequency of the correct segmentation, the frequency of the erroneous segmentation and the length of CW or WS respectively. α, β and γ are coefficients. The optimum coefficients of LEF are decided by the preliminary experiments using Greedy method, α=10, β=1 and γ=5.

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

(3) When 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

Fig. 2 shows an example of a non-segmented sentence. In this example, every character represents a Chinese character, so we use this example to express a general sentence of non-segmented language to present the proposed method. Those words that are not registered in the dictionary are predicted by using Inductive Learning. After the sentences were segmented by known words, which have been registered in the dictionary, the unsegmented part of character string will be used to extract WS. The prediction method is to find the common character string in text. The extraction procedure is carried out as Fig. 3 shows: the extraction of common parts, sift out the common part of the most possibility as a word, the re-extraction of common parts and the extraction of different parts.

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

Figure 2: An example of non-segmented sentence.

2.2.1 Extraction of a Common Part

A common part in non-segmented text is extracted by two steps:

(1) When a character string appears in 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). Here, we use length, frequency and location of S1 in the sentence to sift out it, to get the S1 of the most possible as a word. At this step, we acquired 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 belongs to S1.

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

The extracted S1 at **2.2.1** may still include a common character string. At this situation, the common character string can be re-extracted moreover from the extracted 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 presented as follows:

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

The part of re-extraction is called high dimensional common part and represented by S2 (Segment two). The part of remain is called different part and represented by S3 (Segment three). The S1 is deleted from the dictionary when it is divided into S2 and S3.

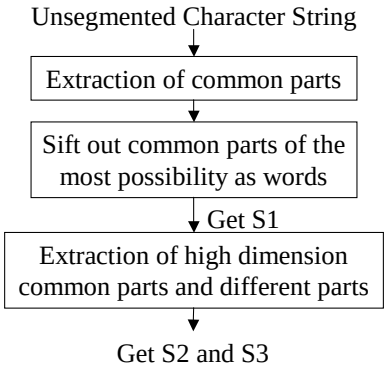

Figure 3: WS extraction procedure

(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 "$\Theta\pi\mu\tau\gamma\beta\pi\alpha\beta\chi\delta$" 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.

The extraction procedure is carried out repeatedly until the new WS can not be extracted and the input can not be segmented.

2.3 Segmentation by a Update Dictionary

The extracted WS are classified to "S1", "S2", and "S3". Those WS that are confirmed as a word by proofreading process are called "CW" (Correct Word). Furthermore, the FR(appearing FRequency), CS(Correct Segmentation frequency), ES(Erroneous Segmentation frequency), LE(LEength) and rank of a word candidate are rigestered simultaneously. Word Segmentation is carried out by the update dictionary as **2.1**.

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 system updates the rank of the registered CW and WS in the dictionary by comparing the corrected results with the segmentation results. 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

Table 1: Experimental results

Fields	Economics	Engineering	Average
words	92,085	70,017	162,102
CSR[%]	87.50	90.80	89.44
ESR[%]	5.40	5.60	5.45
USR[%]	7.10	3.60	5.11

feedback process is described in detail as follows:

(1) For the Correct Segmentation Results:

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

(2) For the Erroneous Segmentation Results:

- If the dictionary does not has the correct words, the system registers the words in the dictionary. In this case, their FRs are 1, their ranks are 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 rank equal CW.

3 Evaluation Experiments

3.1 Experimental Data And Procedure

To evaluate the adaptability of the proposed method for different fields and the effectivity for Chinese word segmentation. We use the Chinese text of two specialized fields from Sinica Corpus[1]: the economics contains 92,085 words and the engineering contains 70,017 words. Total words is 162,102. The economics consists of the text of economic system, economic policy and economic theory. The engineering consists of the text of electronics, communication engineering, machine engineering and nuclear industry.

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

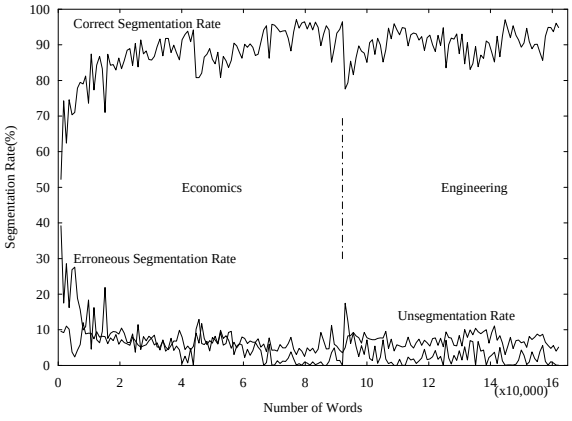

Figure 4: The changes of segmentation rates

In order to confirm the adaptability of proposed method to user, we let the initial dictionary empty. We input a paragraph about hundred words one times and two fields text in turns.

3.2 Experimental Results

The results of experiment are shown in Table 1. Fig. 4 shows the change of CSR, ESR and USR. 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 subtracts the number of correct segmentation and unsegmentation from the number of all words in the input text. To evaluate the experiment result, we use these formulas of CSR (Correct Segmentation Rate), ESR (Erroneous Segmentation Rate) and USR (Unsegmented Rate) as follows:

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

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

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

4 Discussion

4.1 Adaptability To Different Fields

Fig. 4 shows the experimental results of two fields. When the text is changed to different domain, because appearance of some new words of different fields, the correct segmentation rate

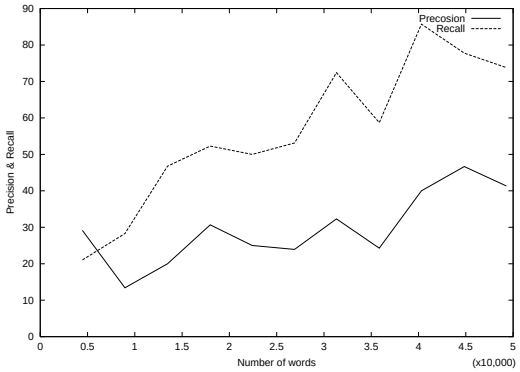

Figure 5: The ability to predict unknown words

is fall down temporary. However with increasing of processed sentence, 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 economics consists of the text of economic system, economic policy and economic theory and so on.

4.2 Evaluation of Ability for Predicting Unknown Words

We use 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 increasing 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.

4.3 Analysis of Erroneous Segmentation

We select 1,000 words from the beginning of the experimental date and the end of the experimental date respectively, to analysis the reason of an erroneous segmentation. At the beginning, ESR that is because of unregistered words is 18.0%, but after 16,000 words are processed, ESR that is because of unregistered words is 0.9%. However ESR that is caused by ambiguity goes on increasing from 1.6% to 7.0%. 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.

5 Conclusion

The experiment results show the prediction ability for unknown words by using Inductive Learning. The experiment results of two fields shown the proposed method can adapt to different fields text. In this paper, the emphasis is to evaluate the adaptivity of the method to different user and fields. About comparison with other existed methods will be done in the future.

The proposed method may be used to computer-aided acquisition of language resource. The experimental results show our proposed method has ability of learning, predictability for unknown words and effectivity for Chinese word segmentation. For the future works, we plan to improve the ability of dealing with segmentation ambiguity, and use this proposed method for Chinese morphological analysis.

References

Kenji Araki, Yoshio Momouchi, and Koji Tochinai. 1995. Evaluation for adaptability of kana-kanji translation of non-segmentation japanese kana sentences using inductive learning. *PACLING-II*, pages 1–7.

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

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

T. Yamasita and Y. Matsumoto. 2000. Journal of natural language processing(in japanese). *Framework for Language Independent Morphological Analysis*, 7(3):39–56.