Industrial Engineering
& Management Systems
Vol 15, No 3, September 2016, pp.224-230
ISSN 1598-7248 | EISSN 2234-6473 |
http://dx.doi.org/10.7232/iems.2016.15.3.224

Language-Independent Word Acquisition Method Using a State-Transition Model

Bin Xu, Naohide Yamagishi
Graduate School of Shonan Institute of Technology, Kanagawa, Japan

Makoto Suzuki*
Department of Information Science, Shonan Institute of Technology, Kanagawa, Japan

Masayuki Goto
Department of Industrial and Management Systems, School of Creative Science and Engineering, Waseda University, Tokyo, Japan

(Received: February 18, 2016 / Revised: July 7, 2016 / Accepted: August 13, 2016)

ABSTRACT

The use of new words, numerous spoken languages, and abbreviations on the Internet is extensive. As such, automatically acquiring words for the purpose of analyzing Internet content is very difficult. In a previous study, we proposed a method for Japanese word segmentation using character N-grams. The previously proposed method is based on a simple state-transition model that is established under the assumption that the input document is described based on four states (denoted as A, B, C, and D) specified beforehand: state A represents words (nouns, verbs, etc.); state B represents statement separators (punctuation marks, conjunctions, etc.); state C represents postpositions (namely, words that follow nouns); and state D represents prepositions (namely, words that precede nouns). According to this state-transition model, based on the states applied to each pseudo-word, we search the document from beginning to end for an accessible pattern. In other words, the process of this transition detects some words during the search. In the present paper, we perform experiments based on the proposed word acquisition algorithm using Japanese and Chinese newspaper articles. These articles were obtained from Japan's Kyoto University and the Chinese People's Daily. The proposed method does not depend on the language structure. If text documents are expressed in Unicode the proposed method can, using the same algorithm, obtain words in Japanese and Chinese, which do not contain spaces between words. Hence, we demonstrate that the proposed method is language independent.

Keywords: Word Segmentation, Character N-gram, Language Independent, State Transition

* Corresponding Author, E-mail: m-suzuki@info.shonan-it.ac.jp

1. INTRODUCTION

In recent years, automatic analysis by computer of text data on the Internet has become increasingly popular. However, in languages that are written without spaces between words, such as Chinese and Japanese, it is necessary to use a specific syntax of the language to perform a morphological analysis. Accurately dividing new words, spoken words, and abbreviations, which appear in large numbers on the Internet, is extremely difficult using conventional morphological analysis techniques because text data include many such words. Therefore, a word segmentation method that can recognize some word information from text data is required. In the present paper, we construct a language-independent word segmentation method. Using this method, spoken words,

new words, and abbreviations can be added automatically to Chinese and Japanese language dictionaries with the same algorithm. Moreover, provided there is a sufficient number of documents, words can be divided with relatively high accuracy.

2. WORD SEGMENTATION

Our method of word segmentation is based on a state-transition model (Yamagishi and Suzuki, 2011). Documents are assumed to be written based on a predetermined state-transition model. Although a number of state-transition models have been proposed, in the present study, we use a simple model, which is shown in Figure 1.

The state-transition model assumes that the input document is described based on four states, namely, states A, B, C, and D, that are specified beforehand. State A represents words (nouns, verbs, etc.), and state B represents punctuation and particle conjunctions (punctuation marks, conjunctions, etc.). State C represents postpositions (namely, words that follow nouns), and state D represents prepositions (namely, words that precede nouns). For example, the initial state is state B. The acceptance state can be any of states A, B, or C.

We initially focus on state B in the state-transition diagram so that phrases can always be divided in front of and behind state B words, because state B is a delimiter or punctuation mark. For example, verbs (state A) and nouns (state A) appear in front of and behind state B words. Finally, postposition (state C) and preposition (state D) words are necessary in order to modify verbs or nouns. The method proposed herein is based on the state-transition model.

Examples of word segmentation are as presented in the following.

An example of a Chinese sentence is shown in Figure 2. Similarly, a Japanese sentence is shown in Figure 3.

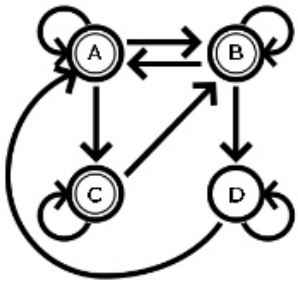

Figure 1. State transition diagram.

我	是	学生	。
A	B	A	B

A 我 | B 是 | A 学生 | B 。

Figure 2. Example of a Chinese sentence.

私	は	学生	です	。
A	B	A	C	B

A 私 | B は | A 学生 | C です | B 。

Figure 3. Example of a Japanese sentence.

3. PREVIOUS METHOD

3.1 Previous Method in Japan

3.1.1 Mochihashi's Method

Mochihashi's method can automatically detect words from character strings in every language without a prepared dictionary (Mochihashi *et al*., 2009). Mochihashi's algorithm can automatically perform word separation in a sentence including spoken words, abbreviations, and new words for all kinds of languages, whereas previous algorithms were not able to deal with such processing. Mochihashi's method assumes the output from a nonparametric Bayesian hierarchical *N*-gram language model of words and characters to be a string. Based on this assumption, "words" are iteratively estimated by a combination of Markov chain Monte Carlo methods and efficient dynamic programming.

3.1.2 Okada and Yamamoto's method

Okada and Yamamoto's approach can automatically obtain candidate words without using a dictionary or supervised data (Okada and Yamamoto, 2013). Their method deals with both known languages and unknown languages. Moreover, it is possible to perform word segmentation even for languages without supervised data. Specifically, Okada and Yamamoto's method calculates the scores of all strings that appear in the text as the product of string length and frequency and adopts segments having high scores as candidate words.

3.1.3 Kaji's method

Kaji suggested that one reason for the low accuracy of word segmentation methods in Japanese is the presence of *katakana*-declinable words (such as *guguru*) (Kaji *et al*., 2009). Therefore, he proposed a word segmentation algorithm that automatically retrieves *katakana*-declinable words from a large-scale network. Kaji's method is as follows:

1) Extract all *katakana* texts from the Internet.
2) Distinguish whether the words in these texts are verbs or adjectives.
3) Extract the *katakana*-declinable words from the verbs and adjectives classified in Step 2.
4) Get rid of all noise words in the post-processing.

3.2 Previous Methods in China

Generally, previously proposed methods that have

been used for Chinese are classified broadly into three approaches:

1) Segmentation is based on a pre-prepared dictionary
2) Segmentation is based on statistics
3) Segmentation is based on artificial intelligence

The proposed method is related to the second approach, namely, segmentation based on statistics, which is performed without a dictionary. In Fu's method, words are considered as combinations of character strings (Fu *et al*., 2002). The more frequently two characters appear adjacent to one another, the more likely these characters are to constitute a word. In this way, processing becomes a purely statistical problem. Furthermore, Xu's method assumes that character strings that appear quite frequently in the same order are more likely to become words (Xu *et al*., 2002). Xu's method performs word segmentation under this assumption. This algorithm can separate new words, spoken words, and abbreviations that occur frequently on the Internet.

4) Neural-network-based representation learning

With the rapid development of neural-network-based representation learning, it has become realistic to learn features automatically. This paper investigated a Chinese word segmentation method based on representation learning (Lai *et al*., 2013). In this previous study, all characters (including numbers and punctuation marks) are classified as belonging to one of four states: the beginning of a word is classified as belonging to state B, the contents of the word are classified as belonging to state M, the end of the word is classified as belonging to state E, and characters that make up words are classified as belonging to state S. This method calculates the relationships between characters by means of a neural network and classifies the characters into four states (The accuracy rate of this method is higher than that of the proposed method, although the proposed method can successfully perform word separation for the Chinese and Japanese languages, neither of which contains spaces between words).

4. PROPOSED METHOD

In the present paper, we propose a new word acquisition method. The proposed method creates a word segmentation dictionary automatically from a certain quantity of text data. Therefore, we use part of the target text data as the learning data to build the word dictionary. In addition, although the text data in the input have delimiters, we do not know what these delimiters are. Moreover, the text data contain words, but we do not know which character strings are words and which are tags (e.g., states A-D) of the character string.

The general flow of the proposed method is as follows.

First, we select the character strings that appear more often as candidate words by using an *N*-gram model. After that, we repeat I) initial selection, II) word segmentation, III) candidate refinement, and IV) localization control in seven phases.

Although each step differs slightly in each of the seven phases, the steps are generally as follows:

1) Initial selection: We initialize the word dictionary if no flag is attached to any of the word candidates. We search the character strings that are above a certain frequency of appearance using 2- to 30-grams. The character strings are reduced one at a time starting from 30-grams. The number of occurrences of relatively large strings is deleted and not used in the subsequent string search. We perform this initial processing only during phase 1.
2) Word segmentation: We separate words using a word dictionary that we configured. We separate the learning documents using the maximum matching method. The character string is registered as a word candidate if the character string is between words, and we register the word candidate as a word if the candidate is not registered in the dictionary. We perform word separation using this dictionary by the subsequent procedures.
3) Candidate refinement: We organize the word dictionary by state-transition diagram. Specifically, we register the candidate word that is more likely to be a word based on the results of the word segmentation. For example, a state B word is often sandwiched between state A words, or a state A word is often sandwiched between state B words. In other words, this processing is intended based on the state-transition model. We repeatedly update the base dictionary based on each transition pattern in the state-transition model until no update is generated.
4) Localization control: We delay convergence in order to avoid over-fitting. At this point, we reset states A and B, and run steps I and II once more. At this point in time, we judge frequently appearing words to belong to state B. Furthermore, we delete words that appear to be incorrectly separated. For example, we delete character strings that do not appear in a word list or those that couple a state A word with a state B word.

Since we cannot trust the dictionary used in the initial selection and word segmentation, we gradually improve the reliability of the dictionary by candidate refinement and localization control. If a reliable dictionary can be obtained, a dictionary-based separation method will be used.

We repeat the aforementioned steps while changing the phase if the result converges to a certain degree. The phases are as follows:

1) Search for words belonging to state AB (steps I-IV)

We separate the state A and B words based only on frequency of appearance using the character *N*-gram.

2) Search for words belonging to state AB+decompose the connection between state AA (steps II-IV)

The character *N*-gram may recognize the character string of state AB as the character string of state A. State AB is defined as a continuous pattern of states A and B. State AA may be a continuous pattern of states A, B, and A. Here, state B may be hidden between two state A strings. We separate these character strings. Moreover, we combine these character strings, because these strings may be fixed phrases that appear frequently.

3) Reset the information of the states thus far (steps II-IV)

We reset the words of state A and state B in the beginning of this step. High-frequency words are classified as belonging to state B in the dictionary.

4) Rerun Phases 1 and 2

We consider the words with very high frequency to be B states and execute phases 1 and 2 again.

5) Search for state ABCD words+decompose and compose the connection between state AA (steps II-IV)

We search state A, B, C, and D words based on the state-transition model. We separate and combine the state AA words. This is the same processing described in phase 2.

6) Search for words belonging to state ABCD+compose the connections between state ABA words (steps II-IV)

The process of "searching for words belonging to state ABCD" is the same as that described in phase 5. On the other hand, we combine the state ABA words. The state ABA words are a continuous pattern of state A, B, and A words, and these character strings may be fixed phrases that appear frequently.

7) Search for words belonging to state ABCD+partially compose the connections between state ABB words or state BBA words (steps II-IV)

The process of "searching for words belonging to state ABCD" is the same as that described in phase 5. On the other hand, we combine the words belonging to states ABB and BBA. State ABB words are a continuous pattern of A, B, and B states. State BBA words are a continuous pattern of state B, B, and A words, and these character strings may be fixed phrases that appear frequently.

The proposed algorithm requires some parameters, the values of which must be set. However, the optimum values of these parameters vary with the number of input documents and the characteristics of the language. Even if we do not provide the optimum values for these parameters, we do not need to set the value strictly because the algorithm only decreases in execution efficiency at a certain range. Moreover, state B is exclusive and preferential to other states. In other words, a word that is classified as belonging to state B cannot belong to any other state.

5. EXPERIMENT

We used Chinese and Japanese corpuses in this experiment. We used the "City University of Hong Kong (CITYU)" Chinese corpus, which consists of 3,147 articles from a period of one month in 1998 from the *People's Daily*, and we used the Japanese corpus "Kyoto University Text Corpus (KUTC)," which was created by Japan's Kyoto University from Mainichi Newspapers Co., Ltd. This corpus consists of 36,767 articles from 1995. We evaluated the final results in Phase 7 based on the following three criteria:

Criteria P (Precision): (the number of positions of "pause between words" in the estimated result that coincide with "pause between words" in the corpus)/(the number of positions that are automatically judged to be "pause between words" by the proposed algorithm)

Criteria R (Recall): (the number of positions of "pause between words" in the estimated result that coincide with "pause between words" in the corpus)/(the number of positions that are described as "pause between words" in the corpus)

Criteria F (F-measure): (2×P×R)/(P+R)

The evaluation results for Chinese and Japanese are presented in Table 1 and Table 2, respectively.

Table 1. Evaluation result (Chinese)

Criteria	Value	Ratio
P	811,501/921,143	88.1%
R	811,501/1,101,129	73.7%
F	(2×P×R)/(P+R)	80.3%

Table 2. Evaluation result (Japanese)

Criteria	Value	Ratio
P	689,170/856,581	90.8%
R	689,170/897,851	72.2%
F	(2×P×R)/(P+R)	80.4%

Table 3. Comparison of experimental results

Model	CITYU	KUTC
NPY(２)	82.4%	62.1%
NPY(３)	81.7%	66.6%
NPY(+)	82.3%	68.2%
proposed method	80.3%	80.4%

[A]十分||[A]关注||[A]最近||[A]一个时期||[A]一些国家

Figure 4. Examples of Chinese results (100).

||[B]の||[A]内閣改造||[B]を||[A]明確||[B]に||[A]否定した||[B]。

Figure 5. Examples of Japanese results (1,000).

Table 3 shows the F-measure. Here, NPY (2) and NPY (3) indicate word bigram and trigram, respectively, and NPY (+) indicates doubled learning data of NPY (3).

Here, we compared Mochihashi's method and the proposed method with the CITYU and KUTC corpuses. The results are shown below.

We set the parameters in the program of the proposed method. These parameters are numerical values that determine the granularity of the word segmentation. The length of a word becomes longer as the value of the parameter becomes larger. We present the results of the evaluation criteria for various values of the parameters in Tables 4 and 5. If we focus on the F-measure, the result of the word segmentation is best when the value of the parameter is 100 in the case of Chinese word segmentation, as shown in Table 4. On the other hand, the best value of the parameter is 1,000 for the case of Japanese word segmentation, as shown in Table 5. Moreover, we present the results of Chinese and Japanese word segmentation in Figures 4 and 5, respectively.

Here, "[A+]" is a word that has an "A" attribute but is composed of multiple words.

On the other hand, we present the results of the Chinese word segmentation when the value of the parameter is 1,000 in Figure 6 and results of the Japanese word segmentation when the value of the parameter is 100 in Figure 7. Moreover, "[N]" denotes "none," which means that the word was not classified because we did not have sufficient information.

In this way, there is a tendency for the length of a word to become longer as the value of the parameter becomes larger.

The number of pseudo-words for each state in the automatically configured word dictionary is indicated as follows. In Tables 6 and 7, the value of a cell is the number of words belonging to state A or state ACD when the value of the parameter is 100. Here, "Del" indicates the number of deleted words or word candidates in the final step. Finally, the value of "None" indicates the number of unrecognized words.

[N]十分关注最近一个时期一些国家

Figure 6. Examples of the Chinese results (1,000).

||[A]内閣改造||[B]を||[A+]明確に否定||[B]し||[B]た||[B]。

Figure 7. Examples of the Japanese results (100).

Table 4. Coincidence rate in the case of changing parameters (Chinese)

Parameters	50	100	200	300	400	500	1,000	1,500
Criteria P	80.46%	88.10%	89.76%	90.38%	91.78%	91.84%	91.89%	91.89%
Criteria R	79.48%	73.70%	69.19%	65.62%	64.30%	63.59%	59.17%	59.17%
Criteria F	79.97%	80.26%	78.14%	76.04%	75.62%	75.15%	71.99%	71.99%

Table 5. Coincidence rate in the case of changing parameters (Japanese)

Parameters	50	100	200	300	400	500	1,000	1,500
Criteria P	79.07%	80.46%	82.53%	81.10%	84.95%	85.76%	90.78%	91.09%
Criteria R	78.54%	76.76%	73.56%	71.94%	70.46%	70.56%	72.17%	70.11%
Criteria F	78.80%	78.57%	77.79%	76.25%	77.03%	77.42%	80.41%	79.23%

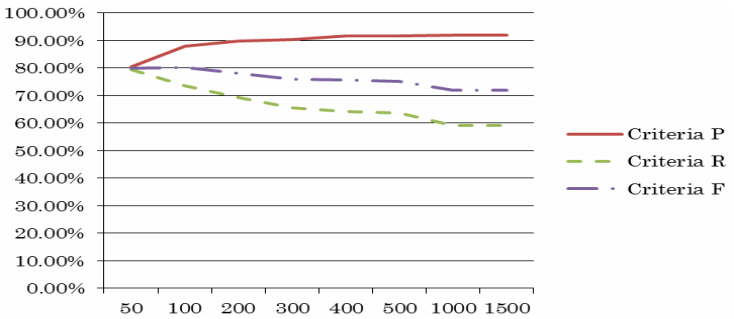

Figure 8. Matching rate in the case of changing parameters (Chinese).

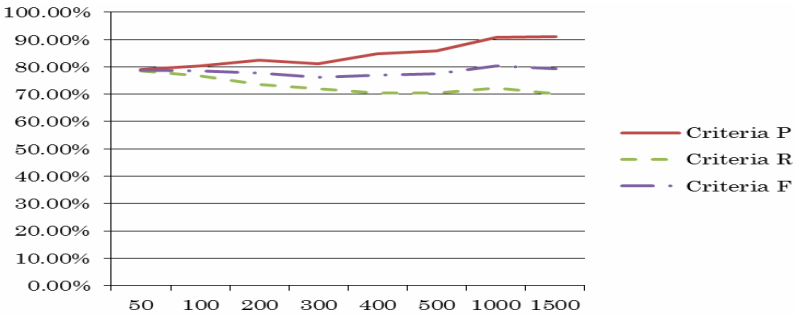

Figure 9. Matching rate in the case of changing parameters (Japanese).

Table 6. State of the Chinese word dictionary when the value of the parameter is 100

A	B	C	D	AC	AD	ACD	Del	None	Sum
21,457	54	9	6	78	70	0	0	162,403	184,077

Table 7. State of the Japanese word dictionary when the value of the parameter is 1,000

A	B	C	D	AC	AD	ACD	Del	None	Sum
25,724	10	21	14	74	62	0	0	113,624	139,529

In addition, we present 31 types of Chinese words from among 54 state B words in the Chinese word dictionary, as follows.

Similarly, we present 10 types of Japanese words from among 10 state B words in the Japanese word dictionary, as follows.

In this way, most of the words that were classified as state B words are appropriate as Chinese and Japanese delimiters, although there were a few incorrect words in these results. We underlined the incorrectly segmented Chinese words in Figures 10 and 12. Similarly, we present 37 types of Chinese words from among 78 state AC words in the Chinese word dictionary, as follows:

，的。在是了和有中为不“”上
要地以来新于等就同与：也而将
从；都

Figure 10. B state (Chinese).

の、。にはをがとでも

Figure 11. B state (Japanese).

突破 上午 贡献 结束 生产力 强调 指出 损失 介绍 救助 水平 考核 座谈会 宣布
今天下午 冲突 问题上 计划 展开 表明 委员会 联盟 两国关系 签署 两岸关系
状态 标准 研究所 模式 适当 之际 意义 设备 准备 副主席 制度 考核

Figure 12. AC state (Chinese).

銀行 会議 委員会 関係者 政治家 だろう
可能性 している システム していない
金融機関 手続き 委員長 労働者 事務所
報告書 首脳会議

Figure 13. AC state (Japanese).

Similarly, we present 17 types of Japanese words from among 74 state AC words in the Japanese word dictionary, as follows.

With few exceptions, these are appropriate words that are considered to have both “nominal usage” and “postpositional usage.” In addition, the size of the dictionary file is only 5.71MB. This is considerably smaller than the size of a dictionary based on a probabilistic model.

6. CONSIDERATION

We compared the experimental results obtained using Mochihashi’s method and those obtained using the proposed method in Section 5. The result obtained using the proposed method for the Chinese corpus was somewhat low, but that for the Japanese corpus was considerably improved.

As shown in Figure 8, the results for the Chinese corpus are exactly the same when the parameter values are 1,000 and 1,500. The proposed method is considered to separate words that have the longest length in each case when the value of the parameter is 1,000 for the Chinese corpus. Accordingly, there is no significance in

setting the parameter to be larger than 1,000. On the other hand, the results for the Japanese corpus are shown in Figure 9. Although the accuracy rates of criteria A and C remain on the same level, that of criteria B tends to decline if the parameter values are set to between 1,000 and 1,500. In addition, the average accuracy rate of criteria C for the Japanese corpus is better than that for the Chinese corpus.

The features common to the Chinese and Japanese results are as follows. The accuracy rates of criteria B and C become smaller and that of criteria A becomes larger as the value of the parameter becomes larger. However, for the Japanese corpus, the rates of change for criteria A and B are lower when the value of the parameter is larger.

7. CONCLUSION

The experiments of the present study reveal that the proposed method facilitates the recognition of Chinese and Japanese words using a corpus without any information about words or to which parts of speech the words belong. The great advantage of the proposed method is that there is no need to set the dependency between individual words. We need to calculate the occurrence probability of the target word in the text data based on the probability method. Therefore, if we obtain new samples, it is necessary to recalculate the probability using the included samples. However, since there are no dependencies between words in the proposed method, we do not need to calculate the probability of the words again, even if we obtain new samples. We can update the dictionary by adding a section containing the words detected from new samples in the dictionary. Hence, once we can construct a dictionary that allows the automatic extraction of words with high accuracy, the proposed method is expected to detect new words or new abbreviations from the new samples quickly. In the future, we intend to investigate the following. First, we will automatically adjust the parameters according to the language or text data. Second, we would like to verify the effectiveness of the proposed method when applied to other languages. Finally, we would like to decrease the number of words that were incorrectly segmented by the proposed method.

ACKNOWLEDGMENT

The present study was supported by JSPS KAKENHI Grant Number 16K01267.

REFERENCES

Fu, S., Yuan, D., Huang, B., and Zhong, Z. (2002), Word extraction without dictionary based on statistics, *Journal of Guangxi Academy of Sciences*, **18**(4) 252-255.

Kaji, N., Fukushima, K., and Kisurekawa, M. (2009), Acquisition of Katakana verbs and adjectives from large Web text, *The IEICE Transactions on Information and Systems*, **J92-D**(3), 293-300.

Lai, S., Xu, L., Chen, Y., Liu, K., and Zhao, J. (2013), Chinese Word Segment Based on Character Representation Learning, *Journal of Chinese Information Processing*, **27**(5).

Mochihashi, D., Yamada, T., and Ueda, N. (2009), *Bayesian unsupervised word segmentation with nested Pitman-Yor language modeling*, ACL, 100-108.

Okada, S. and Yamamoto, K. (2013), *Automatic acquisition of the word-separated units using the occurrence frequency information of the character string*, NLP, 422-425.

Xu, G., Su, X., and Chen, S. (2002), Arithmetic and Application of No Dictionary Cutting Word in Chinese Text Mining, *Journal of Jilin Institute of Technology*, **23**(1), 16-18.

Yamagishi, N. and Suzuki, M. (2011), An unsupervised word acquisition method by adaptation to a state transition model, *Proc. of the 12th Student Paper Presentation of JIMA*, 57-58.