# Extraction of unknown words using the probability of accepting the kanji character sequence as one word

# Hiroyuki Shinnou, Masanori Ikeya

Ibaraki University Dept. of Systems Engineering 4-12-1 Nakanarusawa, Hitachi, Ibaraki, 316-8511, Japan {shinnou,ikeya}@nlp.dse.ibaraki.ac.jp

#### Abstract

In this paper, we propose a method to extract unknown words, which are composed of two or three kanji characters, from Japanese text. Generally the unknown word composed of kanji characters are segmented into other words by the morphological analysis. Moreover, the appearance probability of each segmented word is small. By these features, we can define the measure of accepting two or three kanji character sequence as an unknown word. On the other hand, we can find some segmentation patterns of unknown words. By applying our measure to kanji character sequences which have these patterns, we can extract unknown words. In the experiment, the F-measure for extraction of unknown words composed of two and three kanji characters was about 0.7 and 0.4 respectively. Our method does not need to use the frequency of the word in the training corpus to judge whether its word is the unknown word or not. Therefore, our method has the advantage that low frequent unknown words are extracted.

# 1. Introduction

In this paper, we propose a new method for automatically extracting unknown words, which are composed of two or three kanji characters, from Japanese text. A problem of morphological analysis is the detection of unknown words which are not in the dictionary. In a morphological analysis of a Japanese sentence, an unknown word causes wrong word segmentation, therefore, the detection of unknown words in the text improves the accuracy of morphological analysis. Moreover, the importance of detecting proper nouns in the text for information extraction (IE) has recently been recognized. Most proper nouns such as person names, place names and company names are unknown words, so we need a method to detect unknown words also for IE.

First we need to clarify what unknown words we are concerned with. In particular, we have to take notice of compound words which are handled as one word. For example, let us take the kanji character sequence "自然言語 (natural language)". This sequence is obviously composed of two words "自然 (natural)" and "言語 (language)". If the collocation between these two words is weak, the sequence is judged as a noun phrase composed of two words. In this case, the sequence is not an unknown word. However, if the collocation is strong, this sequence is regarded as a compound word which is treated as one word, that is an idiom chunk. In this case, we can also regard the sequence as an unknown word, if the sequence is not in the dictionary. That is, the difference between a compound word handled as one word and an unknown word is not clear. In this study, we are not concerned with idiom chunks. We are concerned with absolute unknown words like person names, place names, abbreviated words and so on. These words cannot be segmented into other words. Furthermore, we restrict the unknown words to those composed of kanji characters, because many of the other types of unknown words can be detected easily. Consequently, the unknown words which we are concerned with are composed of two or three kanji characters.

To detect unknown words, we introduce the probability of accepting the kanji character sequence as one word. If the character sequence corresponding to a word w is  $\alpha$ , we express w as  $|\alpha|$ . Moreover, we express the probability that the character sequence  $\alpha$  is used as one word as  $P(|\alpha|)$ . We learn  $P(|\alpha|)$  through a training corpus. By using the above probability, we define a measure of the degree to which the character sequence  $\alpha$  is an unknown word.

The unknown word is segmented into other words by the morphological analysis. This can result in more than one pattern. However, the patterns produced by segmentation are limited. That is, we can select word sequences which may be unknown words, by confirming the match between the word sequence and a segmentation pattern. Then we can identify unknown words by applying our proposed measure to selected word sequences.

The most remarkable characteristic of our measure is that we do not use the frequency of  $\alpha$  in the training corpus. Therefore, our method can extract unknown words which have low frequencies.

## 2. Extraction of unknown words

## 2.1. Selection of word sequences

In this paper, we are concerned with the unknown words composed of two or three kanji characters. Furthermore, we make the following assumption:

Assumption If we conduct the morphological analysis for the sentence including the unknown word  $\alpha$  composed of kanji characters, there are word boundaries before and after  $\alpha$ .

This assumption is heuristic, but our experiments had rarely case which violated this heuristic.

With this assumption, if the kanji character sequence ab composed of two kanji characters a and b is an unknown word, the segmentation pattern through the morphological analysis is the following:

$$ab \rightarrow /a/b/$$
 ... Pattern 0 (P0).

Moreover, the kanji character sequence abc is an unknown word, the segmentation pattern through the morphological analysis is one among following three patterns:

abc 
$$\rightarrow$$
 /a/b/c/ ... Pattern 1 (P1)  
abc  $\rightarrow$  /a/bc/ ... Pattern 2 (P2)  
abc  $\rightarrow$  /ab/c/ ... Pattern 3 (P3).

As a result, we can select word sequences which contain unknown words, by confirming the match between the word sequence and the Pattern n.

# 2.2. The probability $P(|\alpha|)$

The probability  $P(|\alpha|)$  that the character sequence  $\alpha$  is used as one word is defined by the following equation:

$$P(|lpha|) = \left\{egin{array}{ll} rac{c(|lpha|)}{c(lpha)} & & if & c(lpha) > 0 \ 1 & & if & c(lpha) = 0 \end{array}
ight.$$

In this equation,  $c(|\alpha|)$  means the frequency of the word  $|\alpha|$  in the training corpus, and  $c(\alpha)$  means the frequency of the character sequence  $\alpha$  in the training corpus.

## 2.3. Measure to identify an unknown word

It seems to be easy to judge whether the character sequence  $\alpha$  is an unknown word or not, because  $P(|\alpha|)$  just means it. However, it is not so easy because we cannot get  $P(|\alpha|)$  from the training corpus and the dictionary if  $\alpha$  is an unknown word. In this paper, we propose the measure of the degree to which  $\alpha$  is an unknown word.

To explain this measure, we take the case that character sequence  $\alpha$  is segmented into two words  $|\beta|$  and  $|\gamma|$ . In this case, the true word segmentation of  $\alpha$  is probably  $|\beta/\gamma|$  or  $|\alpha|$ . The latter means that  $\alpha$  is an unknown word.

The probability that the true word segmentation of  $\alpha$  is  $/\beta/\gamma/$  is almost equal to  $P(|\beta|)P(|\gamma|)$  because the probability that a character sequence is one word is probably independent of words surrounding that sequence. Therefore, the degree to which  $\alpha$  is an unknown word is given by:

$$m(|\alpha|) = 1 - P(|\beta|)P(|\gamma|).$$

By this measure we judge whether  $\alpha$  is an unknown word or not.

However, for the word abc with the Pattern 1, we define m by the following equation:

$$m(|a||b||c|) = max\{1-P(|a|)m(|b||c|), 1-m(|a||b|)P(|c|)\}.$$

# 2.4. The priority order of patterns

First our method picks up word sequences composed of one or two kanji words  $^1$  through the morphological analysis. These picked up word sequences contain unknown words. Next we apply our measure m to word sequences which have the Patterns, that is, P0, P1, P2 or P3. Then we judge whether each word sequence is an unknown word or not.

However, it is ambiguous which pattern the word sequence has, because there are several pattern types. For example, suppose the morphological analysis segments a kanji character sequence abcdefg into /ab/c/d/e/fg/. In this case, /ab/c/, /c/d/e/, /c/d/, /d/e/, and /e/fg/ have the pattern P3, P1, P0, P0 and P2 respectively. Thus, these sequences may be unknown words. If /c/d/e/ is judged as an unknown word, others (/ab/c/,/c/d/,/d/e/ and /e/fg/) are judged not to be unknown words without using our measure.

In this paper, we overcome this problem by setting the priority order of patterns. The order is followings:

$$P1 \rightarrow P0 \rightarrow P2 \rightarrow P3$$
.

In the above example, we first judge whether /c/d/e/ with P1 is an unknown word or not. If we judge that it is an unknown word, we change the segmentation through the morphological analysis, /ab/c/d/e/fg/, to /ab/cde/fg/. Next we judge word sequences with P0, P2 and P3 as this order.

There are cases that the order of word sequences with P0 is ambiguous. In the above example, it is ambiguous which /c/d/ or /d/e/ should be judged first. In this case, we apply the measure to both, and choose one with the bigger measure value.

## 2.5. Exception for numerical expression

We defined the measure of the degree to which the word sequence  $/\beta/\gamma/$  is an unknown word by the following equation:

$$m(|\beta||\gamma|) = 1 - P(|\beta|)P(|\gamma|).$$

If the word  $\beta$  is numerical,  $P(|\beta|)$  is small. As a result,  $m(|\beta||\gamma|)$  is big. However, such word sequence  $|\beta/\gamma|$  is rarely one word. To overcome this problem, we prepare the exception that if the word  $\beta$  is numerical, the word sequence  $|\beta/\gamma|$  is not an unknown word.

# 3. Experiments

## 3.1. Probabilities and threshold values

We conducted the morphological analysis of one year's worth of Mainichi '94 newspaper articles using the Chasen system. By using the morphological analysis result, we calculated  $P(|\alpha|)$  where  $\alpha$  is one or two kanji character sequence.

Furthermore, we set the threshold value of each pattern. The threshold value of P1 was set as 0.90, and other as 0.98. These values were set by investigating a small experiment for a part of the training corpus.

<sup>&</sup>lt;sup>1</sup>In this paper, the n kanji word means the word composed of n kanji characters.

# 3.2. Extraction of unknown words from newspaper articles

To confirm the effectiveness of our proposed method, we picked up 1,000 sentences from the beginning of Mainichi '95 newspaper articles and applied our method to these sentences to extract unknown words from them. In advance, we manually picked up 27 kinds of unknown words from these sentences, which are shown in Table 1.

Table 1: Target unknown words

和魂,硬材,女衆,和慶,彌弌,彩挺, 艷麗,作陶,廣場,星島,羅湖,老子, 祥造,学燈,画期,靖英,普選,仕梅, 万長,深銘,士雄,大三,杉乃井, 三連星,有煙炭,農政通,落馬洲

We show the result of the experiment in Table 2 and extracted words in Table 3. Our method extracted 21 phrases as unknown words, which included 16 true unknown words. That is, this experiment showed that the precision was 0.762, the recall was 0.593 and F-measure was 0.667, which are acceptable. Note that the sign in parentheses in Table 2 shows the pattern number which the extracted word had.

Table 2: Result of experiment

Table 2. Ites are or emperations						
Pattern	Number of	Number of				
	kinds extracted	right answers				
0	16	14				
1	3	2				
2	1	0				
3	1	0				

Table 3: Extracted words

(Right extraction) 画期 (P0), 硬材 (P0), 彩挺 (P0), 作陶 (P0), 仕梅 (P0), 女衆 (P0), 祥造 (P0), 深銘 (P0), 星島 (P0), 普選 (P0), 万县 (P0), 老子 (P0), 和慶 (P0), 和魂 (P0), 杉乃井 (P1), 有煙炭 (P1)

(Wrong extraction) 党論 (P0), 連星 (P0), 張士雄 (P1), 万単位 (P2), 朗読術 (P3)

# 3.3. Frequencies of unknown words in the training corpus

In Table 4, we show the frequencies of target unknown words in the training corpus.

Table 4: Frequencies of unknown words

(extracted unknown words) 画期 (245), 硬材 (0), 彩挺 (0), 作陶 (17), 仕梅 (0), 女衆 (2), 祥造 (1), 深銘 (0), 星島 (15), 普選 (5), 万長 (26), 老子 (3), 和慶 (14), 和魂 (7), 杉乃井 (2), 有煙炭 (0)

(not extracted unknown words) 彌弌(2), 艷麗(0), 廣場(0), 羅湖(0), 学燈(0), 靖英(3), 士雄(3), 大三(158), 三連星(2), 農政通(2), 落馬洲(0)

Table 4 shows that our method is not affected by the frequency of the target unknown word in the training corpus. That is, our method can extract even unknown words which have low frequencies in the training corpus.

# 3.4. Evaluation by word length

In the above experiment, our method for unknown words with P0 that is, unknown words composed of two characters, had that the precision was 0.875, the recall was 0.636, and the F-measure was 0.737. These values are not so bad. On the other hand, our method for unknown words with P1, P2 and P3, that is, unknown words composed of three characters, had that the precision was 0.400, the recall was 0.400, and the F-measure was 0.400. These values are bad.

However, these values cannot evaluate our method for unknown words composed of three characters because these unknown words hardly appear in the test sentences. In this paper, to actually evaluate it, we conducted another experiment to extract unknown words which are forcibly created.

First we picked up words composed of three kanji characters with the frequency 1 from the training corpus and randomly selected 100 words from them. Moreover, we picked up sentences which include selected words. Next we removed selected words from the dictionary of the morphological analysis system Chasen, and rebuilt Chasen. In the rebuilt Chasen, these 100 words are unknown. We applied our method to picked up 100 sentences in order to check whether our method can extract the these 100 words as unknown words or not. We show the result in Table 5.

The sum of numbers of unknown words is not 100 but 86, because 14 words cannot be classified into P1, P2 and P3. The 9 words of them violate our Assumption. The left 5 words are not real words. These 5 words were produced by segmentation errors of Chasen.

In this experiment, the precision of extraction of unknown words composed of three characters was good, but the recall was bad.

Table 5: Extraction of unknown words composed of three characters

Pattern	Number of	Number of extracted words	precision	recall	F-measure
	unknown words	(Number of right extractions)			
P1	18	4 (4)	1.000	0.222	0.363
P2	26	9 (7)	0.778	0.269	0.400
P3	42	12 (10)	0.833	0.238	0.370
SUM	86	25 (21)	0.840	0.244	0.378

#### 4. Discussions

# 4.1. Unknown words composed of three kanji characters

The F-measure of extraction of unknown words composed of two characters was 0.73, which was excellent. However, the case of three characters were not so good. In particular, the recall is bad.

The following is a cause of the poor recall. The unknown word composed of three characters has the pattern P1, P2 or P3. In the case of P2 or P3, the word is composed of two parts. The one is the word composed of two characters  $\alpha$ , and another is the word composed of one character a. A cause of the poor recall is the big  $P(|\alpha|)$ . Moreover, if  $P(|\alpha|)$  is big, P(|a|) tends to be big, too. As a result, we cannot extract the word as the unknown word. The "XX  $\mathcal{F}$  (XX river)", the "XX  $\mathcal{F}$  (XX temple)" and the " $\mathcal{F}$  XX (new XX)" are typical examples. In these examples, most of "XX" are location names and are generally used as one word. So, P(|XX|) is high, and  $P(|\mathcal{F}|)$ ,  $P(|\mathcal{F}|)$  and  $P(|\mathcal{F}|)$  are high, too. As a result, we cannot extract these words as unknown words.

It is difficult to overcome this problem, because this problem is caused by the difficulty of the rigid word definition. In Table 6, we show a part of 100 words which are targets in the second experiment. These words are surely registered as one word, but they can also be regarded as the phrase.

Table 6: Ambiguous words in target unknown words

南町東,	仏照寺,	黄金町,	勝専寺,	大番組,
黒滝川,	石井組,	蓄膿症,	屋島寺,	松村組,
専福寺,	網走川,	海鹿島,	駒形橋,	新茂原,
越前岬,	玄界島,	一合目,	白地図,	学校裏,
真如寺,	清風荘,	放浪癖,	涅槃会,	向洋中,
技術的,	伝久寺,	西桜木,	中島通,	横峰寺,
津島東,	新郡山,	玉出西,	稲荷前,	東八潮

These unknown words are beyond the scope of our method because we are not concerned with them in this paper. To extract them, we must check the strength of the collocation between words. However, the use of the collocation is against our method. To extract unknown words with P2 and P3 is our future work.

#### 4.2. Related research

We can extract unknown words from the text by using the character based method of the automatic word acquisition from the text. The earliest research of this kind was conducted by Nagao and Mori(Nagao and Mori, 1994). They constructed a large N-gram from a corpus by the suffix array method, and extracted character sequences regarded as one word by the diversity of characters before and after the sequence. Moreover, Fung and Wu applied the method acquiring English idiomatic expressions to Chinese text(Fung and Wu, 1994). This method is also character based. Their researches aim not to extract unknown words, but to acquire words. However, viewing their researches from the point of extraction of unknown words, their methods have the defect that the frequency of the unknown word must be high in the training corpus. Furthermore, Mori and Nagao proposed the method to determine whether a character sequence is one word or not and, if so, what its part of speech is by using the distribution of character sequences before and after a part of speech in the training corpus (Mori and Nagao, 1996). However, this method also has the same defect as the above researches.

Nagata extracted unknown words through the morphological analysis system coped with unknown words(Nagata, 1999). First he compute the probability that a character sequence is an unknown word. Next he produce all possible word segmentations including unknown words. By these word segmentations he defined the mean of the frequency that a character sequence appears as one word. We can extract unknown words by regarding the character sequence with the mean beyond a threshold as one word. However in his research, the probability that a character sequence is an unknown word is defined by the product of the word length probability and the word spelling probability. Moreover, the latter is got from the bigram of characters in words. Therefore, this method may not extract unknown words composed of two characters which have low frequencies in the training corpus.

It is clear that we can extract unknown words through the revision of the morphological analysis. The method of Hisamitu and Niwa learns reduction rules to revise the morphological analysis result by the error driven morphological analysis (Hisamitsu and Niwa, 1998). However, they did not evaluate it from the point of the extraction of unknown words. Moreover, Shinnou and Ikeya revised the word segmentation

through the morphological analysis by the character based HMM(Shinnou and Ikeya, 1999). Because most of unknown words are nouns, his method can also extract unknown words from the text. However, it uses frequencies of character sequences in the training corpus to compute the probabilities of output symbols in the HMM. Therefore, their method may not extract unknown words which have low frequencies in the training corpus. In this paper, we set P(|a|) = 1 if the frequency of the character a in the training corpus is 0. This setting does not hinder extracting unknown words.

Furthermore, we can extract unknown words by solving over-segmentations because unknown words are segmented into other words by the morphological analysis. Utiyama solved the over-segmentationed into the word A and the word B by the measure that A and B tend not to segment them(Utiyama, 1999). This research presents the basic idea of our method. However, his proposed measure needs the frequency of the character sequence AB. On the other hand, our method does not need it. Therefore, our method can extract unknown words which have low frequencies in the training corpus.

#### 5. Conclusion

In this paper, we propose a method to extract unknown words composed of two or three kanji characters, from Japanese text.

First, we introduce the probability of accepting the kanji character sequence as one word. By using the probabilities, we define the measure which accepts two or three character sequence as an unknown word. We base this measure on two features. The unknown word must be composed of kanji character which can be segmented into other other discrete words by the morphological analysis and the probability of the appearance of each individual segmented word must be small.

In the experiment, the F-measure for extraction of unknown words which are composed of two and three kanji characters was about 0.7 and 0.4 respectively. Our method does not need to use the frequency of the unknown word in the training corpus to judge whether the word is the unknown word or not. Therefore, our method has the advantage that low frequency unknown words are extracted.

In our future work, we will seek to improve the accuracy of the extraction of unknown words composed of three kanji characters.

## 6. References

Fung, Pascale and Dekai Wu, 1994. Statistical Augmentation of a Chinese Machine-Readable Dictionary. Proceedings of the Second Workshop on Very Large Corpora:69-85.

Hisamitsu, Toru and Yoshiki Niwa, 1998. Postprocessing of Japanese Morphological Analysis Using Transformation Rules and Contextual Information (in Japanese). SIG Notes 98-NL-126, IPSJ. Mori, Shinsuke and Makoto Nagao, 1996. Word extration from Corpora and Its Part-of-Speech Estimation Using Distributional Analysis. Proceedings of the 16th International Conference on Computational Linguistics:1119-1122.

Nagao, Makoto and Shinsuke Mori, 1994. A New Method of N-gram Statistics for Large Number of n and Automatic Extration of Words and Phrases from Large Text Data of Japanese. Proceedings of the 15th International Conference on Computational Linguistics:611-615.

Nagata, Masaki, 1999. Acquisition from Japanese Text Based on Statistical Unknown Word Model and Expected Word Frequency (in Japanese). Journal of Information Processing, 40(9):3373-3386.

Shinnou, Hiroyuki and Masanori Ikeya, 1999. Correction of Word Segmentation Errors through Character-based HMM. The 4th Conference of Pacific Association for Computational Linguistics (PACLING '99):131-136.

Utiyama, Masao, 1999. Statistical Measure for Detecting Over-Segmentations in Results of Japanese Morphological Analysis (in Japanese). Journal of Natural Language Processing, 6(7):3-28.