

Segmentation of Compound Nouns using Composite Mutual Information

Kwangseob Shim
School of Computer Science and Engineering
Sungshin Women's University
Seoul 136-742, KOREA
shim@cs.sungshin.ac.kr

Abstract

In Korean, a compound noun may be freely formed with or without spaces between simple nouns. The flexible word formation rule of Korean raises a serious problem in processing compound nouns with computers, in particular, in searching a dictionary with the compound noun as a search key. This paper describes a corpus-based method for segmenting a compound noun into simple nouns. Segmentation is performed with the composite of four types of mutual information, each calculated from two adjacent syllables. Experiments with a test data set of 1,458 compound nouns show that the correct segmentation rate of the proposed method is 98.3%. A text corpus of 1.1 million words (4.9 million syllables) is used for the calculation of composite mutual information.

1 Introduction

Compound nouns are formed by combining any number of simple nouns in sequence. They carry more specific contextual information than simple nouns and thus they have been considered very important in natural language processing. The problem with compound nouns is that we cannot register all of them in an on-line dictionary since the number of compound nouns is too big.

A reasonable solution to the problem is to segment a compound noun into simple nouns. This task may be trivial for languages such as English, where simple nouns are usually separated by spaces. The situation is worse, however, in Korean where a compound noun may be freely formed with or without spaces between the simple nouns (Park et al. 1996). Moreover, arbitrarily long compound nouns are possible in Korean. For example, compound nouns of 5 simple nouns or longer are not rare in real texts. Compound nouns of that length are usually composed of more than 10 syllables.

Although a Korean compound noun can be formed with or without spaces between the simple nouns, we assume in this paper that there is no space in a compound noun. This assumption is reasonable because only the portion of a given compound noun, if it is still a compound noun and there are no spaces in it, need to be considered for segmentation. With this assumption, the segmentation problem is exponential because there are 2^n segmentation possibilities for a compound noun of n syllables (Yosiyuki et al. 1994). Similar problems were reported to occur in Chinese, Japanese and German, where no spaces are placed between simple nouns (Hisamitsu and Nitta 1996; Pachunke et al. 1992; Wong and

Chan 1996; Yosiyuki et al. 1994).

In this paper, we propose a corpus-based method of segmenting Korean compound nouns into simple nouns. Segmentation is performed using the composite of four types of mutual information, each calculated from two adjacent syllables.

The proposed method may be applied to a word identification problem — to segment a sentence into words, if all the words in the sentences are concatenated without any space between the words. The problem has been considered important in Chinese natural language processing systems (Chen and Liu 1992; Sproat et al. 1994).

2 Mutual Information

The mutual information, originally defined by Fano (Fano 1961), has been used in natural language processing to capture the relationship between two specific words (Church and Mercer 1993). The mutual information $I(x, y)$ is defined to be

$$I(x, y) \equiv \log_2 \frac{P(x, y)}{P(x) \cdot P(y)}, \quad (1)$$

where x and y are two words that appear in a specific language. Mutual information compares the joint probability of observing x and y together with the word probabilities of observing x and y independently. If there is a strong association between x and y , then $P(x, y)$ will be much larger than $P(x)P(y)$, and consequently $I(x, y) \gg 0$. Otherwise, $I(x, y)$ will be almost equal to or much less than 0 (Church and Hanks 1990).

With a corpus of size N , the word probabilities $P(x)$ and $P(y)$ are estimated by $\frac{f(x)}{N}$ and $\frac{f(y)}{N}$. Similarly, the joint probability $P(x, y)$ is estimated by

$\frac{f(x,y)}{N}$. Therefore, the estimate of mutual information is defined to be

$$I(x,y) \approx \log_2 \frac{N \cdot f(x,y)}{f(x) \cdot f(y)}. \quad (2)$$

As the corpus size N increases, the estimate will get more accurate.

There were some efforts to use word co-occurrence information in solving the segmentation problem (Hisamitsu and Nitta 1996). Because of the large number of different words, however, this approach raised a data sparseness problem — some word co-occurrence information is missing that is needed to solve a problem. A possible solution to the data sparseness problem is to collect the word co-occurrence information from a much larger corpus.

Another approach, we propose, is to use syllable distribution information. For a given compound noun with no spaces in it, segmentation may be paraphrased as deciding whether a space should be inserted or not between any two adjacent syllables. Mutual information of two syllables is used for that decision. The next section describes the basic segmentation method using mutual information of two adjacent syllables.

3 Basic Segmentation Method using Mutual Information of Syllables

Before describing the basic method, let me introduce the spelling rule of Korean. Although Korean is a phonographic language, our spelling rule is particular and systematic. Thus, there is no hyphenation problem in Korean.

We have 14 consonants and 10 vowels. A syllable always consists of a leading consonant, a vowel and an optional trailing consonant. All vowels are variants of a horizontal line or a vertical bar. Variants of a horizontal line are placed below the leading consonant, whereas variants of a vertical bar are placed to the right of the consonant. The trailing consonant is always placed below the vowel. Double consonants and diphthongs follow the same spelling rule. For example, “함수” is a word (means a *function*) of two syllables. The first syllable “함” consists of a leading consonant “ㅎ”, a vowel “ㅏ”, which is a variant of the vertical bar “|”, and a trailing consonant “ㅁ”. The second syllable “수” consists of a leading consonant “ㅅ” and a vowel “ㅜ”, which is a variant of the horizontal line “_”. Since a syllable is encoded in two bytes, *computationally* it is an easy task to segment a word into syllables.

Returning back to the segmentation issue, let me describe how to calculate mutual information of two adjacent syllables. Assume a compound noun with spaces between the simple nouns. For example, consider a compound noun of two simple nouns $s_1 \cdots s_m$ and $s_{m+1} \cdots s_n$

$$s_1 \cdots s_m \quad s_{m+1} \cdots s_n,$$

where s_i represents a syllable. For all $i = 1, 2, \dots, n$, the frequency $f(s_i)$ increases by one while the joint frequencies remain unchanged except $f(s_m, s_{m+1})$ that increases by one. In this way, the joint frequencies are counted for all the compound nouns in a corpus. Then, the mutual information $I(s_i, s_{i+1})$ of two adjacent syllables s_i and s_{i+1} is calculated by Eq.2.

The basic segmentation method is very simple. Consider a compound noun that has no spaces in it.

$$s_1 \cdots s_i s_{i+1} \cdots s_n$$

For all $i = 1, 2, \dots, n-1$, insert a space between two adjacent syllables s_i and s_{i+1} if mutual information $I(s_i, s_{i+1})$ exceeds a threshold. The threshold is determined empirically.

The number of different syllables used in a language is more limited than the number of different words used in that language. In Korean, almost 10,000 different syllables are possible according to the spelling rule mentioned above. However, only part of them are used in real texts. For example, just 2,000 or less different syllables are used in a Korean dictionary which has about 100,000 entries. Therefore, the data sparseness problem is less serious when we are using the mutual information of *syllables*. Reliable statistics on syllables will be obtained from a text corpus of reasonable size.

However, there is a problem with the basic method described in this section. The problem occurs because the mutual information of syllables carries less information than mutual information of words. An experiment showed that the segmentation accuracy was as low as 77% when the basic segmentation method was applied. If more than two syllables are considered simultaneously, the accuracy will rise as a matter of course. However, as we move from bigram to trigram, a much bigger corpus is required in order to get reliable statistics. So, that is not our direction to go. Instead, the composite mutual information is proposed to relieve the problem while maintaining the advantages of using mutual information of syllables over using word co-occurrence information.

4 Composite Mutual Information

4.1 Four Types of Mutual Information

While keeping the basic segmentation method, we slightly modify the calculation method of mutual information of two adjacent syllables. With the modification, there are four types of mutual information.

- $I_p(s_i, s_{i+1})$

As the mutual information $I_p(s_i, s_{i+1})$ increases, the possibility of segmenting between s_i and s_{i+1} also increases. In this sense, the mutual information is called *positive*. Positive mutual information is exactly the one that we have used in the previous section.

- $I_n(s_i, s_{i+1})$

As the mutual information $I_n(s_i, s_{i+1})$ increases, the possibility of segmenting between s_i and s_{i+1} decreases. In this sense, the mutual information is called *negative*. Negative mutual information is acquired in a similar way we described in the previous section. Consider a compound noun with a space between s_m and s_{m+1}

$$s_1 \cdots s_m \ s_{m+1} \cdots s_n,$$

where s_i represents a syllable. The above compound noun will be used below in describing $I_h(s_i, s_{i+1})$ and $I_t(s_i, s_{i+1})$. For all $i = 1, 2, \dots, n-1$, a joint frequency $f_n(s_i, s_{i+1})$ increases by one except $f_n(s_m, s_{m+1})$ that remains unchanged. The new joint frequency $f_n(s_i, s_{i+1})$ is used in the calculation of negative mutual information $I_n(s_i, s_{i+1})$.

- $I_h(s_i, s_{i+1})$

As the mutual information $I_h(s_i, s_{i+1})$ increases, the possibility of segmenting before s_i increases. In this sense, the mutual information is called *head*. For all $i = 1, 2, \dots, n-1$, the joint frequency $f_h(s_i, s_{i+1})$ remains unchanged except $f_h(s_{m+1}, s_{m+2})$ that increases by one. The new joint frequency $f_h(s_i, s_{i+1})$ is used in the calculation of head mutual information $I_h(s_i, s_{i+1})$.

- $I_t(s_i, s_{i+1})$

As the mutual information $I_t(s_i, s_{i+1})$ increases, the possibility of segmenting after s_{i+1} increases. In this sense, the mutual information is called *tail*. For all $i = 1, 2, \dots, n-1$, the joint frequency $f_t(s_i, s_{i+1})$ remains unchanged except $f_t(s_{m-1}, s_m)$ that increases by one. The new joint frequency $f_t(s_i, s_{i+1})$ is used in calculation of tail mutual information $I_t(s_i, s_{i+1})$.

4.2 Definition of Composite Mutual Information

Each type of mutual information defined above carries morphological information, so that it may be used in segmentation. However, the segmentation accuracy is not good when each type of mutual information is used independently. For example, the accuracy was merely 77% when segmentation was performed only with the positive mutual information.

The segmentation accuracy will be maximized when the four types of mutual information are considered simultaneously. It is difficult, however, to develop a segmentation algorithm using the four types of mutual information as a vector, *i.e.*, $\langle I_p, I_n, I_h, I_t \rangle$. Thus, we propose that the four types of mutual information be composed into one scalar value.

The composition is achieved by adding each type of mutual information multiplied by an appropriate weight. The following equation defines the composite mutual information.

$$\begin{aligned} I_c(s_i, s_{i+1}) &= w_p \cdot I_p(s_i, s_{i+1}) \\ &+ w_n \cdot I_n(s_i, s_{i+1}) \\ &+ w_h \cdot I_h(s_{i+1}, s_{i+2}) \\ &+ w_t \cdot I_t(s_{i-1}, s_i) \end{aligned}$$

where w_p, w_n, w_h and w_t are weights.

By definition, as the negative mutual information $I_n(s_i, s_{i+1})$ increases, the possibility of segmenting between s_i and s_{i+1} decreases. The other types of mutual information has the same characteristics in that the increase in mutual information indicates the increased possibility of segmenting between s_i and s_{i+1} . Thus, w_n should be negative whereas w_p, w_h and w_t are always positive. Weights are determined empirically. Through the paper, the following weights are assumed.

$$w_p = 0.6, w_n = -0.5, w_h = 0.6, w_t = 0.6$$

Composite mutual information $I_c(s_i, s_{i+1})$ may be used in deciding whether a space should be inserted or not between s_i and s_{i+1} . Note that the composite mutual information, although the arity is 2, carries a co-occurrence information of four syllables s_{i-1}, s_i, s_{i+1} and s_{i+2} .

5 The Algorithm

As mentioned in Section 1, we are dealing with compound nouns that has no spaces in it. Consider a compound noun of n syllables

$$s_1 s_2 \cdots s_m s_{m+1} \cdots s_n.$$

Let $I_c(s_m, s_{m+1})$ be the highest composite mutual information of all $I_c(s_i, s_{i+1})$ where $i = 1, 2, \dots, n-1$. If $I_c(s_m, s_{m+1})$ exceeds a threshold, for example, a space is inserted between s_m and s_{m+1} . In this way, the given compound noun is segmented into two. If this segmentation is correct, both segments may be either a simple noun or a compound noun, shorter than the original. In the latter case, segmentation will be continued in the same way as we did above. In fact, the segmentation procedure is recursively applied until the given compound noun is completely broken down into simple nouns.

In some cases, segmentation may be incorrect since the statistics is not always perfect. Usually, the statistics carries noises. Nevertheless, we may assume that most of the cases the segmentation is performed correctly if the threshold is determined in such a way that the segmentation accuracy is maintained within an allowable level.

Now consider the case in which the highest composite mutual information does not exceed the threshold. This means that the possibility of incorrect segmentation is relatively high. In this case, the

recursive application of the segmentation procedure described above may not be practical. Thus, segmentation is performed in a different way, but still guided by the composite mutual information.

Let $I_c(s_m, s_{m+1})$ be the highest composite mutual information and do not exceed a threshold. The given compound noun is temporarily segmented between s_m and s_{m+1} . There are four cases with this temporary segmentation.

- (1) Both segments $s_1 s_2 \cdots s_m$ and $s_{m+1} s_{m+2} \cdots s_n$ are simple nouns.
- (2) The left segment $s_1 s_2 \cdots s_m$ is the only simple noun.
- (3) The right segment $s_{m+1} s_{m+2} \cdots s_n$ is the only simple noun.
- (4) Segmentation occurs in the middle of a simple noun, and thus both segments are neither a simple noun nor a compound noun.

In the case of (1), the temporary segmentation becomes permanent and the procedure stops because there is nothing to be segmented further.

In the case of (2), if the right segment is a compound noun, the segmentation may be considered correct. Otherwise, the right segment would be merely a meaningless sequence of syllables. The latter case occurs when the left segment is subsumed by a simple noun longer than that. Consider the following examples.

- (2.1) 국제원자력기구 \longrightarrow 국제 원자력기구
(2.2) 무방향성전기강판 \longrightarrow 무방 향성전기강판

In (2.1), a Korean compound noun “국제원자력기구” is segmented into “국제” and “원자력기구” in which the left segment is a simple noun and the right segment is a compound noun. The right segment will be further segmented later.

In (2.2), a compound noun “무방향성전기강판” is segmented “무방” and “향성전기강판” in which “무방” itself is a simple noun whereas “향성전기강판” is nothing but a meaningless sequence of syllables. Here, the left segment “무방” is subsumed by a simple noun “무방향성”. In fact, the given compound noun should be segmented into a simple noun “무방향성” and a compound noun “전기강판”.

The problem with the case (2) is that it is difficult to differentiate (2.1) from (2.2). The problem is dealt with by finding the longest simple noun and then re-segmenting the given compound noun right after the newly found longest simple noun. The longest simple noun will be found by moving *rightwards*, one syllable at a time, from the point where the initial segmentation occurs. In (2.1), since the simple noun “국제” itself is the longest simple noun, re-segmentation is not necessary. In (2.2), since the longest simple noun “무방향성” is found, re-segmentation occurs right after the newly found simple noun “무방향성”.

The case (3) will be the same with the case (2) except that *right* and *left* should be exchanged.

In the case of (4), we re-apply the above procedure to the given compound noun, not with the highest composite mutual information, but with the next highest composite mutual information. This procedure is repeated until one of the first three cases out of the four is met.

The formal description of our segmentation algorithm is shown in Figure 1. In this figure, a function `segmentUsingCompositeMI(l, h)` segments a compound noun $s_l s_{l+1} \cdots s_h$ into simple nouns. A function `lookupDictionary(l, h)` is a boolean function that returns *true* if $s_l s_{l+1} \cdots s_h$ is registered in the dictionary as a simple noun. Otherwise, it returns *false*. A function `argmax(l, h, ith)` returns an integer m if $I_c(s_m, s_{m+1})$ is the ith highest composite mutual information of all $I_c(s_i, s_{i+1})$ where $i = l, l+1, \dots, h-1$. A function `insertSpace(m)` inserts a space between s_m and s_{m+1} . Finally, a function `findLongestNoun($l, h, m, direction$)` finds the longest simple noun, starting at s_m , moving to the direction given by *direction*. The *direction* may be either *leftwards* or *rightwards*. If the *direction* is *leftwards*, the function returns an integer $m' > l$ such that $s_{m'+1} s_{m'+2} \cdots s_m s_{m+1} \cdots s_h$ is the longest simple noun. Note that the longest simple noun ends in the initial simple noun $s_{m+1} s_{m+2} \cdots s_h$. If the *direction* is *rightwards*, the function returns an integer $m' < h$ such that $s_l s_{l+1} \cdots s_m s_{m+1} \cdots s_{m'}$ is the longest simple noun. Now, the longest simple noun begins with the initial simple noun $s_l s_{l+1} \cdots s_m$.

```

procedure segmentUsingCompositeMI( $l, h$ )
begin
  if (lookupDictionary( $l, h$ )) then return;
   $m = \text{argmax}(l, h, 1)$ ;
  if ( $I_c(s_m, s_{m+1}) > \text{threshold}$ ) then
    segmentUsingCompositeMI( $l, m$ );
    segmentUsingCompositeMI( $m+1, h$ );
    insertSpace( $m$ );
  else
    segmentUsingLongestNoun( $l, h, 1$ );
  endif
end

```

```

procedure segmentUsingLongestNoun( $l, h, ith$ )
begin
  if ( $ith > h - l + 1$ ) then return;
   $m = \text{argmax}(l, h, ith)$ ;
  RHSfound = lookupDictionary( $m+1, h$ );
  LHSfound = lookupDictionary( $l, m$ );
  if (RHSfound AND LHSfound) then
    insertSpace( $m$ );
  else if (RHSfound) then
     $m = \text{findLongestNoun}(l, h, m, \text{leftwards})$ ;
    segmentUsingCompositeMI( $l, m$ );
    insertSpace( $m$ );
  else if (LHSfound) then

```

```

m = findLongestNoun(l, h, m, rightwards);
segmentUsingCompositeMI(m + 1, h);
insertSpace(m);
else
segmentUsingLongestNoun(l, h, ith + 1);
endif
end

```

Figure 1: The Algorithm

6 Dictionary Format for Speed-up

Consider again the compound noun of n syllables

$$s_1 s_2 \cdots s_m s_{m+1} \cdots s_n.$$

Composite mutual information $I_c(s_i, s_{i+1})$ is calculated for all $i = 1, 2, \dots, n - 1$. For this calculation, a *null* syllable s_0 is assumed to the left of s_1 . Similarly, a *null* syllable s_{n+1} is assumed to the right of s_n . The joint frequency may be provided by a dictionary of the following format.

$$s_a s_b \quad f_p(s_a, s_b) \quad f_n(s_a, s_b) \quad f_h(s_a, s_b) \quad f_t(s_a, s_b)$$

where s_a and s_b represent a syllable, respectively. $s_a s_b$ is used as a search key. With this format, the dictionary will be accessed $(n + 1)$ times in order to calculate $I_c(s_i, s_{i+1})$ for all $i = 1, 2, \dots, n - 1$.

According to the segmentation algorithm shown in Figure 1, a simple noun dictionary is used to determine whether the segmented noun is a simple noun or not. Since dictionary access takes much time, we propose to integrate the joint frequency dictionary and the simple noun dictionary, so that the total number of dictionary access may be minimized. The integrated dictionary format is as follows.

$$s_a s_b \quad f_p(s_a, s_b) \quad f_n(s_a, s_b) \quad f_h(s_a, s_b) \quad f_t(s_a, s_b)$$

L_1 A list of simple nouns that begin with $s_a s_b$.
 L_2 A list of simple nouns that end in $s_a s_b$.

When the dictionary entries are accessed for the calculation of composite mutual information, the accompanying noun lists are loaded into main memory. When we converted our simple noun dictionary into the proposed format, the maximal length of the list was 46 and the average length was 2.62. This means that the length of the lists is not arbitrarily long.

L_1 is used by `lookupDictionary()` to decide whether the given word is a simple noun or not. L_1 is also used by `findLongestNoun()` to find the longest noun moving *rightwards*. L_2 is used by `findLongestNoun()` to find the longest noun moving *leftwards*.

7 Examples

In this section, three examples are provided to show how and in what order compound nouns are segmented into simple nouns. First, consider a korean compound noun “동아시아태평양담담차관보”. The following table shows the composite mutual information sorted in decreasing order.

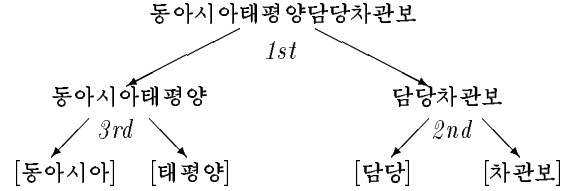
rank	$I_c(x, y)$		rank	$I_c(x, y)$	
1	$I_c(\text{양, 담})$	11.36	7	$I_c(\text{태, 평})$	-14.45
2	$I_c(\text{담, 차})$	5.87	8	$I_c(\text{담, 담})$	-16.33
3	$I_c(\text{아, 태})$	-3.31	9	$I_c(\text{차, 관})$	-16.41
4	$I_c(\text{평, 양})$	-11.61	10	$I_c(\text{관, 보})$	-17.51
5	$I_c(\text{시, 아})$	-13.57	11	$I_c(\text{동, 아})$	-19.31
6	$I_c(\text{아, 시})$	-13.58			

A threshold is assumed to be 5.0. By this value, the first two exceed the threshold. According to the segmentation algorithm shown in Figure 1, spaces will be inserted between “양” and “담”, and “담” and “차”. As a result, the given compound noun will be segmented as shown below.

동아시아태평양 담담 차관보.

Now consider the next highest composite mutual information $I_c(\text{아, 태})$. It does not exceed the threshold. According to our algorithm, a compound noun “동아시아태평양” is temporarily segmented into “동아시아” and “태평양”. Since both segments are simple nouns, the segmentation becomes permanent and the procedure stops.

The following is a graphical representation of the sequence in which segmentation is performed. In this tree, the bracketed word represents a simple noun. As understood from this figure, the given compound noun is segmented, without fail, just in three attempts and this shows how effectively the composite mutual information guides the segmentation.



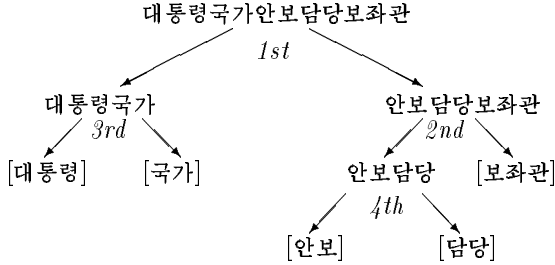
Next, consider a compound noun “대통령국가안보담담보좌관”. The following table shows the composite mutual information sorted in decreasing order.

rank	$I_c(x, y)$		rank	$I_c(x, y)$	
1	$I_c(\text{가, 안})$	13.34	7	$I_c(\text{국, 가})$	-12.66
2	$I_c(\text{담, 보})$	6.00	8	$I_c(\text{담, 담})$	-13.46
3	$I_c(\text{평, 국})$	4.73	9	$I_c(\text{통, 평})$	-17.00
4	$I_c(\text{보, 담})$	1.18	10	$I_c(\text{좌, 관})$	-23.21
5	$I_c(\text{안, 보})$	-3.99	11	$I_c(\text{대, 통})$	-23.62
6	$I_c(\text{보, 좌})$	-10.36			

In this table, the highest composite mutual information $I_c(\text{가, 안})$ exceeds the threshold. Thus, the given compound noun is segmented into “대통령국가” and “안보담담보좌관”. Since both segments are compound, segmentation continues with the next highest composite mutual information $I_c(\text{담, 보})$. Since it exceeds the threshold, “안보담담보좌관” is segmented into a compound noun “안보담담” and a simple noun “보좌관”.

The third highest composite mutual information $I_c(\text{령,국})$ does not exceed the threshold. According to the algorithm, “대통령국가” is temporarily segmented into “대통령” and “국가”. Since both segments are simple nouns, the temporary segmentation becomes permanent. In the same way, “안보담당” is segmented into “안보” and “담당”, by using the fourth highest composite mutual information $I_c(\text{보,담})$.

The following tree shows the sequence in which segmentation is performed.

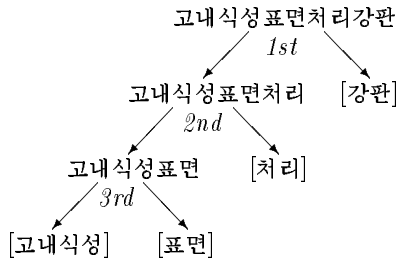


Finally, consider a compound noun “고내식성표면처리강판”. The following table shows the composite mutual information sorted in decreasing order.

rank	$I_c(x, y)$	rank	$I_c(x, y)$
1	$I_c(\text{리,강})$ 16.38	6	$I_c(\text{식,성})$ -14.10
2	$I_c(\text{면,처})$ 15.72	7	$I_c(\text{표,면})$ -16.05
3	$I_c(\text{성,표})$ 10.86	8	$I_c(\text{처,리})$ -17.64
4	$I_c(\text{고,내})$ -9.01	9	$I_c(\text{강,판})$ -30.22
5	$I_c(\text{내,식})$ -13.90		

In this table, we have $I_c(\text{리,강})$, $I_c(\text{면,처})$ and $I_c(\text{성,표})$ that exceed the threshold. By using each composite mutual information in that order, the given compound noun is segmented into four simple nouns.

The following tree shows the sequence in which segmentation is performed.



8 Experiments

A text corpus of 1.1 million words (4.9 million syllables) was used for the calculation of composite mutual information. The corpus contains compound nouns which have spaces between simple nouns. The weights we mentioned in Section 4.2 were used in the calculation of composite mutual information.

A test data set was prepared to evaluate the proposed segmentation algorithm. It consists of 1,458

compound nouns (4,322 simple nouns), which has been randomly extracted from 1994 *Dong-A Ilbo*, one of the Korean major daily newspapers. Since the text corpus was based on scientific documents, the test data set is completely independent of the corpus.

The following table shows the length distribution of the test data set. The shortest one is 4 syllables long and the longest one is 15 syllables long.

length of compound nouns (in syllables)	test data set size (in words)
4	281
5	271
6	208
7	181
8	165
9	132
10	123
11 ~ 15	97
total	1,458

If we consider the segmentation as a problem of deciding whether a space should be inserted or not between two adjacent syllables, the accuracy may be calculated as follows.

$$\frac{\text{the number of correctly segmented points}}{\text{the length of a compound noun} - 1} \quad (3)$$

where the length is given as the number of syllables. For example, consider the Korean compound noun “대통령국가안보담당보좌관”. It consists of 12 syllables. As we have seen in Section 7, the compound noun should be segmented as follows.

대통령 국가 안보 담당 보좌관

To show how the accuracy is calculated, assume that the given compound noun is segmented as follows.

대통령 국가안 보담당 보좌관

↑ ↑ ↑ ↑ ↑ ↑ ↑ ↑

where the uparrow represents a correctly segmented point, and the underlined word represents a simple noun. The word which is not underlined represents a meaningless sequence of syllables.

According to Eq.3, the accuracy is given as $\frac{8}{12-1} = 72.7\%$. This calculation method may be used in such languages as Chinese and Japanese, in which words are not separated by spaces.

Since the method is based on syllables, however, it is not appropriate for such languages as Korean where words are separated by spaces. Moreover, the accuracy tends to be overestimated. More realistic accuracy is calculated as follows.

$$\frac{\text{number of simple nouns correctly segmented}}{\text{the length of a compound noun}} \quad (4)$$

where the length is given as the number of simple nouns, assuming that the given compound noun is correctly segmented.

Consider again the example shown above. Since the length of the given compound noun is 5 and there are two simple nouns that are correctly segmented, using Eq.4, the accuracy is re-calculated as $\frac{2}{5} = 40\%$. Compare this value with the accuracy calculated with Eq.3.

The following table shows the relationship between the accuracy and the threshold. Here, the accuracy was calculated with Eq.4. When the threshold was 8.0, the accuracy was 98.3%.

threshold	accuracy (%)
5.0	97.1
6.0	97.4
7.0	98.0
8.0	98.3

The following table shows the relationship between the accuracy and the length of compound nouns. The length was given as the number of syllables.

length (in syllable)	accuracy (%)	
	threshold = 5.0	threshold = 8.0
4	100	100
5	98.2	98.3
6	98.4	99.5
7	97.0	98.7
8	96.1	97.1
9	96.8	98.6
10	95.7	97.0
11 ~ 15	94.0	97.1

9 Conclusions

In this paper, we proposed a corpus-based compound noun segmentation method. The method uses composite mutual information, calculated from syllable co-occurrence information. Since the number of different syllables in a language is much smaller than that of different words in the same language, syllable co-occurrence information, and thus the composite mutual information, may be easily acquired from a text corpus of manageable size.

Experiments were performed with 1,458 compound nouns (4,322 simple words), which was randomly extracted from a Korean newspaper. The shortest compound noun was 4 syllables long and the longest one was 15 syllables long. Composite mutual information was calculated from a text corpus of 1.1 million words (4.9 million syllables). The corpus size is relatively small. Nevertheless, the proposed segmentation method showed a good performance. The maximal accuracy was 98.3%. This showed how effectively the syllable co-occurrence information, and thus composite mutual information, guided the segmentation.

We applied the proposed method to a more complex problem — to segment a sentence into words, assuming that all the words in the given sentence are concatenated without any space between the words. Experiments showed the accuracy was around 90%.

References

- [Chen and Liu 1992] Keh-Jiann Chen and Shing-Huan Liu, “Word identification for Mandarin Chinese Sentence,” *Proceedings of the 14th International Conference on Computational Linguistics*, pp.101-107, 1992.
- [Church and Hanks 1990] Kenneth Church and Patrick Hanks, “Word Association Norms, Mutual Information, and Lexicography,” *Computational Linguistics*, Vol.16, No.1, pp.22-29, 1990.
- [Church and Mercer 1993] Kenneth Church and Robert L. Mercer, “Introduction to the Special Issue on Computational Linguistics Using Large Corpora,” *Computational Linguistics*, Vol.19, No.1, pp.1-24, 1993.
- [Fano 1961] R. Fano, *Transmission of Information: A Statistical Theory of Communications*, MIT Press, Cambridge, 1961.
- [Hisamitsu and Nitta 1996] Toru Hisamitsu and Yoshihiko Nitta, “Analysis of Japanese Compound Nouns by Direct Text Scanning,” *Proceedings of the 16th International Conference on Computational Linguistics*, pp.550-555, 1996.
- [Pachunke et al. 1992] T. Pachunke, O. Mertineit, K. Wothke and R. Schmidt, “Broad Coverage Automatic Morphological Segmentation of German Words,” *Proceedings of the 14th Conference on Computational Linguistics*, pp.1218-1222, 1992.
- [Park et al. 1996] Hyouk R. Park, Young S. Han, Kang H. Lee, Key-Sun Choi, “A Probabilistic Approach to Compound Noun Indexing in Korean Texts,” *Proceedings of the 16th International Conference on Computational Linguistics*, vol.1, pp.514-518, 1996.
- [Sproat et al. 1994] Richard Sproat, Chilin Shih, William Gale, and Nancy Chang, “A Stochastic Finite-state Word-segmentation Algorithm for Chinese,” *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*, pp.66-73, 1994.
- [Wong and Chan 1996] Pak-kwong Wong and Chorkin Chan, “Chinese Word Segmentation based on Maximum Matching and Word Binding Force,” *Proceedings of the 16th International Conference on Computational Linguistics*, pp.200-203, 1996.

[Yosiyuki et al. 1994] Kobayasi Yosiyuki, Tokunaga Takenobu, Tanaka Hozumi, “Analysis of Japanese Compound Nouns using Collocational Information,” *Proceedings of the 15th International Conference on Computational Linguistics*, pp.865-869, 1994.