# Retrieving Collocations From Korean Text 

Seonho Kim, Zooil Yang, Mansuk Song<br>\{pobi, zooil, mssong\}@december.yonsei.ac.kr<br>Dept. of Computer Science,<br>Yonsei University, Seoul, Korea

Jung-Ho Ahn<br>jungho@math.yonsei.ac.kr<br>Dept. of Mathematics, Yonsei University, Seoul, Korea


#### Abstract

This paper describes a statistical methodology for automatically retrieving collocations from POS tagged Korean text using interrupted bigrams. The free order of Korean makes it hard to identify collocations. We devised four statistics, 'frequency', 'randomness', 'condensation', and 'correlation' to account for the more flexible word order properties of Korean collocations. We extracted meaningful bigrams using an evalnation function and extended the bigrams to n-gram collocations by generating equivalence sets, $\alpha$-covers. We view a modeling problem for n-gram collocations as that for clustering of cohesive words.


## 1 Introduction

There have been many theoretical and applied works related to collocations. A rapidly growing availability of copora has attracted interests in statistical methods for automatically extracting collocations from textual corpora. However, it is not easy to identify the central tendencies of collocation distribution and the borderlines of criteria are often fuzzy because the expressions can be of arbitrary lengths in a large variety of forms. Getting reliable collocation patterns is particularly difficult in Korean which allows arguments to scamble so freely. This paper presents a statistical method using 'interrupted bigrams' for automatically retrieving collocations and idiomatic expressions from Koream text. We suggest several statistics to account, for the more flexible word order.

If the distribution of a random sample is unknown, we often try to make inferences about its properties described by suitably defined measures. For the properties of arbitrary collocation distribution, four measure statistics: 'high frequency', 'condensation', 'randomness', and
'correlation' were devised.
Given a morpheme, our system begins by retrieving the frequency distributions of all bigrams within window and then meaningful bigrams are extracted. We produce $\alpha$-covers to extend them into n-gram collocations ${ }^{1}$.
According to the definition of Kjellmer and Cowie, a fossilized phrase is a sequence, where the occurrence of one word almost predicts the rest of the phrase and one word predicts a very limited number of words in a semi-fossilized phrase (Kjellmer, 1995) (Cowie, 1981). However, in both fossilized and semi-fossilized types there is a high degree of cohesion among the members of the phrases (Kjellmer, 1995). We consider the cohesions as $\alpha$-covers that are obtained by applying a fuzzy compatibility relation, which satisfies symmetry and reflexivity, to meaningful bigrams. Namely, n-gram collocations could be interpreted as equivalent sets of the meaningful bigrams through partitioning. Here, $\alpha$-covers mean the clustered sets of the meaningful bigrams.

## 2 Related Works

In determining properties of collocations, most of corpus-based approaches accepted that the words of a collocation have a particular statistical distribution(Cruse, 1986). Although previous approaches have shown good results in retrieving collocations and many properties have been identified, they depend heavily on the frequency factor.
(Choueka et al., 1983) proposed an algorithm for retrieving only uninterrupted collocations, ${ }^{2}$

[^0]since they assumed that a collocation is a sequence of adjacent words that frequently appear together. (Church and Hanks, 1989) defincel a collocation as a pair of correlated words anid used mutual information to evaluate such lexical correlations of word pairs of length two. They retrieved interrupted word pairs, as well as uninterrupted word pairs. (Haruno et al., 1996) ronstructed collocations by combining adjacent n-grams with high value of mutual information. (Breidt, 1993)'s study was motivated by the fact that mutual information could not give realistic figures to low frequencies and used $t$-score for a significance test for V-N combinations.

Martin noted that a span of 5 words on left and right sides captures $95 \%$ of significant collocations in English (Martin, 1983). Based on this assumption, (Smadja, 1993) stored all bigrams of words along with their relative position, $\mathrm{p}(-5$ $\leq p \leq 5)$. He evaluated the lexical strength of a word pair using ' $Z$-score' and the variance of its position distribution using 'spread'. He defined a collocation as an arbitary, domain dependent, recurrent, and cohesive lexical cluster.
(Nagao and Mori, 1994) developed an algorithm for calculating adjacent n-grams to an arbitrary large number of $n$. However, it was hard to find an efficient $n$ and a lot of fragments were obtained. In Korean, statistics based on adjacent n-grams is not sufficient to capture various types of collocations. (Shimohata et al., 1997) employed entropy value to filter out fragments of the adjacent n-gram model. They evaluated disorderness with the distribution of adjacent words preceding and following a string. The strings with a high value of entropy were accepted as collocations. This disorderness is efficient to eliminate fragments but can not handle interrupted collocations. In general, previous studies on collocations have dealt with restricted types and depend on filtering measures in a lexically point of view.

## 3 Input Format

In this section, we discuss an input form relevant to Korean language structure and linguistic contents which would work well on an effi-
minterrupted collocation is a sequence of words. To avoid confusion of terms, we call a sequence of two words as a 'adjacent bigram' and a sequence of n words as a a.djacent n-gram'.
cient statistics. Korean is one of agglutinative languages as well as a propositional language. An elementary node being called as 'eojeol' is generally composed of a content word and function words. Namely, a word in English corresponds to a couple of morphemes in Korean.

A key feature of Korean is that function words, such as propositions, endings, copula, auxiliary verbs, and particles, are highly developed as independent morphemes, while they are represented as word order or inflections in English. Functional morphemes determine grammatical relations, tense, modal, and aspect. In Korean, there are lots of multiple function words in a rigid forms. They can be viewed as collocations. For this reason, our system is designed at the morphological level. A set of twelve part of speech tags, \{ N, J, V, P, D, E, $\mathrm{T}, \mathrm{O}, \mathrm{C}, \mathrm{A}, \mathrm{S}, \mathrm{X}\}^{3}$ was considered.
Another feature is a free word order. Since the words of a collocation appear in text with the flexible ways, sufficient samples are required to compute accurate probabilities. We allow positional information to vary by using an interrupted bigram model.

The basic input can be represented in (1). An object k means a pair of morphemes ( $\mathrm{m}_{i}, \mathrm{~m}_{k}$ ) and $\mathrm{m}_{k}$ corresponds to one of all possible morphemes, being able to co-occur with $\mathrm{m}_{i}$. A variable j indicates the j -th position. $\mathrm{X}_{i j}$ denotes the frequency of $m_{k}$ that occurs at the $j$-th position before $\mathrm{m}_{i}$.

$$
X_{i}=\left(\begin{array}{cccc}
X_{11} & X_{12} & \cdots & X_{110}  \tag{1}\\
X_{21} & X_{22} & \cdots & X_{210} \\
\vdots & \vdots & \vdots & \vdots \\
X_{n 1} & X_{n 2} & \cdots & X_{n 10}
\end{array}\right)
$$

Given a predicate morpheme as a base morpheme, the range of window is from -1 to -10 . This distance constraint is for the characteristic of SOV language. If a bigram includes an adverb morpheme, a larger window, from -20 to 10 is used because the components often appear widely separated from each other on text. In other cases, we considered the range from -5 to +5 . This distant constraints are for an efficient statistics.

An input data is transformed to a property matrix, $\mathrm{T}\left(\mathrm{X}_{i}\right)$ as (2) that is a two dimensional

[^1]| $\left(\mathrm{m}_{1}, \mathrm{~m}_{4}\right)$ |  | syntactic relation | preferring position |
| :---: | :---: | :---: | :---: |
| (마시,앖이) | (drink,minch) | VD | 1 |
| (마시,너우) | (drink, too) | VD | 2 |
| (마시,맑) | (drink, clean) | VJ | 4 |
| (아시, 매일) | (drink, everyday) | VD | 3 |
| (아시, 끞이) | (drink, boil) | VV | 2 |
| (마시,안) | (drink, not) | VD | 1 |
| (마시, 란께) | (drink, together) | VD | 3 |
| (마시, 조금씩) | (drink, a little) | VD | 1 |
| (마시,반) | (drink, take) | VV | 2 |
| (아시,조긍씩) | (drink, a little) | VD | 3 |

Figure 1: meaningful bigrams of 마시(drink) by Xtract
array of $k$ objects, $k=1,2, \ldots, n$, on four variables, $V_{\text {Frequency }}, \mathrm{V}_{\text {Condensation }}, \mathrm{V}_{\text {Randomness }}$, and $\mathrm{V}_{\text {Correlation }}$.

$$
T\left(X_{i}\right)=\left(\begin{array}{cccc}
V_{1 F} & V_{1 C} & V_{1 R} & V_{1 C R}  \tag{2}\\
V_{2 F} & V_{2 C} & V_{2 R} & V_{2 C R} \\
\vdots & \vdots & \vdots & \vdots \\
V_{n F} & V_{n C} & V_{n R} & V_{n C R}
\end{array}\right)
$$

To continue explanations, we begin by mentioning the 'Xtract' tool by Smadja (Smadja, 1993). Our input form was designed in a similar manner with 'Xtract'. Smadja assumed that the components of a collocation should appear together in a relatively rigid way because of syntactic constraint. Namely, a bigram pair ( $m_{i}$, $m_{k}$ ), where $m_{k}$ occurs at one(or several) specific position around $m_{i}$, would be a meaningful bigrams for collocations. The rigid word order is related with the variance of frequency distribution of $\left(m_{i}, m_{k}\right)$. 'Xtract' extracted the pairs whose variances are over a threshold and pulled out the interesting positions of them by standardization of the frequency distributions. Unfortumately, the approach for English has several limitations to work ${ }^{4}$ on Korean structure for the following reasons:

1. For free order languages such as Korean, words are widely distributed in text, so that positional variance affects the overfiltering of tiseful bigrams. Figure 1 shows that there is no pair which contains randomly distributed morphems such as function words or nouns. This indicates that very few pairs were produced when 'Xtract' is applied to Korean.

[^2]2. Suppose that a meaning bigram, $\left(\mathrm{m}_{i}, \mathrm{~m}_{k}\right)$ prefers a position $\mathrm{p}_{j}$. Then, the number of concordances for condition probability $P\left(m_{i}, m_{k} \mid p_{j}\right)$ would be small, specially in a free order language. As shown in Table 1, the model produced a lot of long meaningless n -grams when compiling into n -grams. The precision value of Korean version of Xtract was estimated to be $40.4 \%$.
3. The eliminated bigrams by the previous stage can appear again in n-gram collocations. When compiling, the model only keeps the words occupying the position with a probability greater than a given threshold from the concordances of $\left(\mathrm{m}_{i}\right.$, $\mathrm{m}_{k}, \mathrm{p}_{j}$ ). As one might imagine, the first stage could be useless.

As stated above, in Korean, the effect of position on collocations needs to be treated in some complex ways. Korean collocations can be divided into four types: 'idiom' ${ }^{5}$, 'semantic collocation' ${ }^{6}$, 'syntactic collocation' ${ }^{7}$, and 'morphological collocation' ${ }^{8}$. Idioms and morphological collocations appear on text in a rigid way and word order but others do in the flexible ways. From a consideration of these more flexible collocations, we adopt an interrupted bigram model and suggest several statistics that consist with characteristics of Korean.

## 4 Algorithm

This section describes how properties are represented as numerical values and how meaningful objects are retrieved. In the first stage, we extract meaningful interrupted bigrams based on four properties. Next, the meaningful bigrams are extended into n -gram collocations using a $\alpha$-compatibility relation.

It empirically showed that a Weibull distribution (3) provides a close approximation of frequency distribution of bigrams.

$$
\begin{equation*}
F(x)=1-e^{-\alpha x^{\beta}} \quad 0<x<\infty \text { where } \alpha>0, \beta>0 \tag{3}
\end{equation*}
$$

[^3]| fred | n－grams | dist | eval |
| :---: | :---: | :---: | :---: |
| 11 | 모두（everyone）．Noun 를（object case）．．$\langle\langle$ 받 $\rangle\rangle$ 어（take）$\langle\langle$ 마시 $\rangle\rangle$（drink） | －2 | $\bigcirc$ |
| 12 |  | －1 | $\bigcirc$ |
| $\underline{2}$ | 두（two）다리（legs）에（at（location case））조금（a little）힘 를 주고（strain）서（stand）어서（with～ing） 친구（friend）와의（with）얘기를 계속하（talk over）면서（～ing）〈（조금씩》）（a little）커피（coffee）를（object case） $\langle\langle$ 마시 》〉（Jrink）는（modifying ending）꿈（dream）．． | －3 | X |
| 1 | ．．．두（two）시간（hours）를（object case）〈（함께 $\rangle\rangle($ together ）술（alcohol）를（object case）$\langle\langle$ 마시 $\rangle\rangle(d r i n k)$ 은（modifying ending）$ᄀ$（he）．． | －3 | 0 |
| 4 | 골라（cola）를（object case）〈＜조금씩）〉（a little）〈＜마시〉＞（drink）어 가며（ $\sim$ ing ） | －1 | （） |
| 1 | 나（I）도（also）백림（baeklim（location））에서（at）〈〈매일〉〉（everyday）비루（tears）를（object case） $\langle\langle$ 마시 $\rangle\rangle$（dink）고있 었（was～ing）기 에（because）．．． | －3 | X |
| 4 | 나（1）여기（here）특히（specially）새벽（dawn）에（at）《〈맑〉〉（fresh）은（modifying ending）물（water）도（also）〈\｛마시〉〉（drink）．다는 ．것（that）부당하（be unsuitable），가장（most）．고 하 면서（says）운동（exercise）잘（well） | －4 | X |
| 6 | ．．Verb ．．Noun 를（object case）$\langle\langle$ 녀무〉》（too）많이（much）$\langle\langle$ 마시 》〉（drink）은（modifying ending） | －2 | 0 |
| \％ | Noun Noun ．．．Noun 속에서（in）$\langle\langle$ 꿇이 $\rangle\rangle$（boil）어 $\langle\langle$ 마시 $\rangle\rangle$（drink）는 Noun | －2 | x |
| 8 | ．．는데도（even though）．$\langle\langle$ 안 $\rangle\rangle$（not）$\langle\langle$ 마시 $\rangle$（drink）고（and）Noun Noun．Verb Ending Verb | －1 | O |

Table 1：$n$－grams of 마시（drink）by Xtract（freq：freq of sentences）

This，there are a lot of pairs with low frequency which interrupt to get reliable statistic．We climinated such pairs using median $m$ that is a value such that $P\{X \geq m\} \geq 1 / 2$ to a fre－ quency distribution $F$ ．If median is less than 3 ， we took the value 3 as a median．

Any quantity that depends on not any un－ known parameters of population distribution but only the sample is called a statistic．We regarded four statistics relating to properties of collocations as variables．Before the further ex－ planation，consider $S_{m_{i}}$ ，a sample space of $m_{i}$ as Table 2 whose cardinality $\left|S_{m_{i}}\right|$ is n ．Let one ob－ ject be $\left(\mathrm{m}_{i}, \mathrm{~m}_{k}\right)$ and its frequency distribution be $f_{i k_{1}}, f_{i k_{2}}, \cdots, f_{i k_{10}}$ and $f_{i k_{+}}$be $\sum_{p=1}^{10} f_{i k_{p}}$ ． Suppose that $\operatorname{POS}\left(\mathrm{m}_{i}\right)$ is＇ J ＇and $\operatorname{POS}\left(\mathrm{m}_{k}\right)$ is ＇P＇．

## 4．1 Properties

The properties which we considered are primar－ ily concerned with the frequency and positional iuformations of word pairs．As we have em－ phasized，the correlation between position and collocation is very complicated in Korean．

According to Breidt，MI or T－score thresh－ olds work satisfactory as a filter for extraction of collocations，but filtered out at least half of the actual collocations（Breidt，1993）．Gener－ ally，assumed properties could not fully account for collocations．Therefore，in order to reduce a loss of information，the combination of observed variables would be better than filtering．We de－ fined four variables for properties of collocations
as follows．
1．$V_{f}$ ：According to Benson＇s definition，a colloca－ tion is a recurrent word combination（Ben－ son et al．，1986）．We agree with this view that a word pair of high frequency would be served as a collocation． $\mathrm{V}_{f}$ statistic of an object $\left(\mathrm{m}_{i}, \mathrm{~m}_{k}\right)$ ，is represented as（4）． Here，standardization demands attention． The mean and standard deviation are cal－ culated in the＇JP＇set which the object be－ longs to．

$$
\begin{align*}
& V_{f}=\frac{f_{i k_{+}}-\bar{f}_{i J P}}{\sigma_{i J P}}, \\
& \text { where } \bar{f}_{i_{J P}}=\frac{\sum_{l=1}^{n} f_{i l+}}{n}=\frac{f_{i++} \mid J P}{n}  \tag{4}\\
& \sigma_{i_{J P}}=\sqrt{\frac{\sum_{l=1}^{n}\left(f_{i l_{+}}-\bar{f}_{i_{J P}}\right)^{2}}{n}}
\end{align*}
$$

2．$V_{c}$ ：Intuitively，two words that prefer spe－ cific positions must be related with each other．We seeked to recapture the idea with the flexibility of word order．For this，the concept of convergence on each position was employed．In a free or－ der language，a meaningful pair can oc－ cur in text either with two distance or three distance．Let＇s consider two in－ put vectors $\mathrm{x},(0,1,0,0,0,1,0,0,1,0)$ and y ， $(0,0,0,1,1,1,0,0,0,0)$ ．They have the same variance but y would be more meaning－ ful than $x$ ，because $y$ can be interpreted as $(0,0,0,0,3,0,0,0,0,0)$ within the free or－ der framework．Therefore，a spatial mask

| word pair | POS pair | total frequency | variable(position) distribution |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| ( $\mathrm{m}_{i}, \mathrm{~m}_{1}$ ) | (J,P) | $f_{i 1+}$ | $f_{i 11}$ | $f_{i 12}$ | $\ldots$ | $f_{i 110}$ |
| $\left(\mathrm{m}_{i}, \mathrm{~m}_{2}\right)$ | (J,P) | $f_{i 2+}$ | $f_{i 21}$ | $f_{i 22}$ | $\ldots$ | $f_{i 210}$ |
|  | $\vdots$ | : | 引 | $\vdots$ | ; | : |
| $\left(\mathrm{m}_{i}, \mathrm{~m}_{k}\right)$ | (J,P) | $f_{i k_{+}}$ | $f_{i k_{1}}$ | $f_{i k_{2}}$ | $\ldots$ | $f_{i k_{10}}$ |
|  | : | : | : | : | : |  |
| $\left(\mathrm{m}_{i}, \mathrm{~m}_{n}\right)$ | (J,P) | $\mathrm{fin}_{+}$ | $f_{\text {in }}$ | $f_{\text {in }}$ | $\ldots$ | $f_{i 1_{10}}$ |
| total |  | $f_{i++}{ }^{\text {J }}$ P | $f_{i+1\}}{ }^{\text {d }}$ | $f_{i+2 \mid J P}$ | ... | $f_{i+10} \mid J P$ |

Table 2: all combinations of $m_{i}$ under a JP relation
( $1 / 2,1,1 / 2$ ) was devised for convergence on each position. The calculation of condensation value $m_{i k_{p}}$ at p -th position is:
$m_{i k_{p}}= \begin{cases}\frac{4 f_{i k_{1}}+3 f_{i k_{2}}+f_{i k_{3}}}{} & p=1 \\ \frac{f_{i k_{p-1}}+2 f_{i k_{p}}+f_{i k_{p}+1}}{2} & p=2 \ldots 9 \\ \frac{f_{i k_{8}}+3 f_{i k_{9}}+4 f_{i k_{1} 0}}{4} & p=10\end{cases}$
The $m_{i k_{p}}$ is computed by neighborhoods that are located in the border of the p-thl position. The $\max _{p} \frac{m_{i k_{p}}}{f_{i k_{+}}}$is likely to represent the condensation of $\left(m_{i}\right.$, $m_{k}$ ) but it is still deficient. Intuitively, ( $0,1,1,1,0,3,2,0,0,0$ ) would be less condensed than $(0,0,3,0,0,3,2,0,0,0)$. Therefore, $n^{\prime}$ was designed for a penalty factor.

$$
\begin{equation*}
V_{r:}=\max _{p=1,2, \ldots, 10} \frac{m_{i k_{p}}}{\sqrt{n^{\prime}} f_{i k_{+}}} \tag{5}
\end{equation*}
$$

$n^{\prime}$ is the number of m , such that $f_{i k_{m}} \neq 0$ for $0 \leq m \leq 10$, and it is a reverse proportion to the condensation. Square root was used for preventing the excessive influence of $n$ '.
3. $V_{r}$ : We were motivated by the idea that if a pair is randomly distributed in terms of position. then it would not be meaningful. Especially in the case of function words, they are likely to be randomly distributed over a given morpheme but distributions of meaningful pairs are not random, as shown in Figure 3. A typical method for the check of randomness is to measure how far the given distribution is away from a uniform distribution. In (6), $\bar{f}_{i k}$ means the expected number of $\left(\mathrm{m}_{i}, \mathrm{~m}_{k}\right)$ at each position on the assumption that the pair randomly occurs
at the position. $\frac{\left|f_{i k_{p}}-\bar{f}_{i k}\right|}{\bar{f}_{i k}}$ can be viewed as an error rate at each position $p$ based on the assumption. The big difference between the expected number and the actual observed frequency means that the distribution is not random. One might think that this concept is the same with one of variance. However, note the denominator. This calculation is somewhat better than variance which depends on frequency.

$$
\begin{equation*}
V_{r}=\sum_{p=1}^{10}\left(\frac{f_{i k_{p}}-\bar{f}_{i k}}{\bar{f}_{i k}}\right)^{2} \tag{6}
\end{equation*}
$$

4. $V_{c r}$ : To become a meaningful bigram, a pair should be syntactically valid. We viewed that if the frequency distribution of a pair keeps the overall frequency distribution of the POS relation set which the pair belongs to, then the pair would be syntactically valid. To verify this idea, we depict the overall frequency distributions in some POS relations in Figure 2. It shows the frequency distributions of pairs which are composed of postposition and predicate morpheme. It is quite interesting that all objects have the similar form of frequency distribution. They have sharp peaks at the first and third position. Clearly, this illustrates that a postposition has a high probability of appearing at the first and third position before a predicate. We can conclude from this that pairs keeping the overall frequency structure would be syntactically valid. We used correlation coefficient for the structural similarity. In the case of a pair $m_{i k}$, the correlation value between $\left(f_{i k_{1}}, f_{i k_{2}}, \ldots, f_{i k_{10}}\right)$ and $\left(f_{i+1 \mid I P}\right.$. $\left.f_{i+2 \mid J P}, \ldots, f_{i+{ }_{10} \mid J P}\right)$ is evaluated. Let x


Figure 2: Frequency distributions of pairs with 'JP' or 'VP' POS relation
and $y$ be two vectors whose components are mean corrected, $x_{i}-\bar{x}$ for $\mathbf{x}, y_{i}-\bar{y}$ for $y$. The correlation between two variables is straightforward, if x and y is standardized through dividing each of their elements by the standard deviations, $\sigma_{\mathrm{x}}$ and $\sigma_{\mathrm{y}}$, respectively. Let $\mathrm{x}^{*}$ be $\mathrm{x} / \sigma_{\mathrm{x}}$ and $\mathrm{y}^{*}$ be $\mathrm{y} / \sigma_{\mathrm{y}}$, then the correlation between x and $\mathrm{y}, V_{c r}$ can be represented as follows.

$$
\begin{align*}
& \mathrm{x}^{\prime}=\left(f_{i k_{1}}, f_{i k_{2}}, \ldots, f_{i k_{10}}\right) \\
& \mathrm{y}^{\prime}=\left(f_{i+1}, f_{i+1}, f_{i+2} \mid J P, \ldots, f_{i+{ }_{10} \mid J P}\right)  \tag{7}\\
& V_{c r}=\frac{\mathbf{x}^{+} \mathrm{y}^{*}}{10}
\end{align*}
$$

The ranks of bigrams by four measures is summarized in Figure 3. It tells that each of the measures comes up with our expectation.

### 4.2 Evaluation Function

in this section, we analyze the correlations of four measures we defined and explain how to make an evaluation function for extracting meaningful bigrams. Table 3 shows the values of correlations which exist in the given measures: $V_{f}, V_{r}, V_{c}, V_{c r}$. This explains that the defined measures have redundant parts. We can say that if a measure has the high values of

|  | $V_{f}$ | $V_{c}$ | $V_{r}$ | $V_{c r}$ |
| :---: | :---: | :---: | :---: | :---: |
| $V_{f}$ | 1.0 |  |  |  |
| $V_{c}$ | -0.495 | 1.0 |  |  |
| $V_{r}$ | -0.203 | 0.506 | 1.0 |  |
| $V_{c r}$ | 0.252 | -0.278 | -0.002 | 1.0 |

Table 3: correlations between factors
correlations between others, then it has a redundant part to be eliminated. Since we don't know what factors are effective in determining useful bigram, the concept of weights is more reliable than filtering. We constructed an evaluation function, which reflects the correlations between the measures.

First of all, we standardized four measures. Standardization gives an effect on adjustment of value range according to its variability. The degree of relationship between measure 1 and measure 2 can be obtained by $C_{\text {measurel, measure } 2}$ which is $\{\operatorname{corr}$ (measure 1, measure 2$)\}^{+}$, where $x^{+}=x$ if $\mathrm{x} \geq 0, x^{+}=0$ otherwise. The evaluation function is concerned with the degrees of relationships of measures.

$$
\begin{equation*}
f\left(V_{f}, V_{r}, V_{c}, V_{c r}\right)=V_{f}+\phi_{r} V_{r}+\phi_{c} V_{c}+\phi_{c r} V_{c r} \tag{8}
\end{equation*}
$$

Here, the constant $\mathbf{a}(\approx 0.845)$ is for a compensation coefficient. The minimum value of $\phi_{r}$, $\phi_{c}$ and $\phi_{c r}$ is $1 / 3$ respectively, where $C_{V_{f}, V_{r}}=$ $C_{V_{f}, V_{c}}=C_{V_{f}, V_{c r}}=0$ and all correlations of $V_{r}, V_{c}$, and $V_{c r}=1$. On the contray, the maxmum value of $\phi_{r}, \phi_{c}$, and $\phi_{c r}$ is 1 respectively, where $C_{V_{f}, V_{r}}=C_{V_{f}, V_{c}}=C_{V_{f}, V_{c r}}=0$ and all correlations of $V_{r}, \dot{V}_{c}, V_{c r}=0$. In other words, as the coefficients $\phi_{r}, \phi_{c}$, and $\phi_{c r}$ get closer to 1 , the correlations between measures reduce.

As shown in (8) and (9), we agree that $V_{f}$ is a primaryl factor of collocations. Each coefficient $\phi$ indicates how much the property is reflected in evaluation. For example, in the case of $\phi_{r}, \mathbf{a} \frac{C_{v_{r}, v_{c}}}{2}$ is a portion which is related with the property of condensation within randomness. Therefore, $1-\mathbf{a} \frac{C_{v r}, v_{c}}{2}$ corresponds to the remainder, when subtracting this portion from randomness.

The threshold for evaluation was set by testing. When the value for threshold was 0.5 , good results were obtained but in noun morphems, a high value over 0.9 was required. The pairs are selected as meaningful bigrams whose values of the evaluation function are greater than the threshold.

### 4.3 Extending to n-grams

The selected meaningful bigrams from the previous step are extended into n-gram collocations. At the final step, the longest ones among all cr-covers are obtained as n-gram collocations by eliminating substrings. Here, n-gram collocations mean interrupted collocations as well as n-character strings.

We regarded cohesive clusters of the meaningful bigrams as $n$-gram collocations on the assumption that members in a collocation have a high degree of cohesion (Kjellmer, 1995). To find cohesive clusters, a fuzzy compatibility relation $R$ is applied. $R$ on $X \times X$, where $X$ is the set of all meaningful bigrams which contain a base morpheme $m_{i}$, means a cohesive relation and partitions of set $X$ obtained by $R$ correspond to n-gram collocations. To say shortly, our problem has shifted to clustering of a set $X$. A reason to employ the concept of fuzzy is that
equivalence sets defined by the relation may be more desirable.

A fuzzy compatability relation $R(X, X)$ is represented as a matrix by a membership function. The membership function of a fuzzy set $\mathrm{A} \in$ $X$ is denoted by $\mu_{A}: X \rightarrow[0,1]$ and maps elements of a given set $X$ into real numbers in $[0,1]$. These two membership functions $\mu_{A}$ were used to define the cohesive relation as follows.

$$
\begin{gather*}
\mathbf{p}(\mathbf{x})=\frac{|\mathbf{x} \cap \mathbf{y}|}{|\mathbf{x}|}, \mathbf{p}(\mathbf{y})=\frac{|\mathbf{x} \cap \mathbf{y}|}{|\mathbf{y}|} \\
\mu_{\mathbf{A}}(\mathbf{x}, \mathbf{y})= \begin{cases}\mathbf{D}(\mathbf{p}(\mathbf{x})| | \mathbf{p}(\mathbf{y})) & =\mathbf{p}(\mathbf{x})\left(\log \frac{1}{\mathbf{p}(\mathbf{y})}-\log \frac{1}{\mathbf{p}(\mathbf{x})}\right) \\
& \text { if } \mathbf{p}(\mathbf{x}) \leq \mathbf{p}(\mathbf{y}) \\
\mathbf{D}(\mathbf{p}(\mathbf{y}) \| \mathbf{p}(\mathbf{x})) & =\mathbf{p}(\mathbf{y})\left(\log \frac{1}{\mathbf{p}(\mathbf{x})}-\log \frac{1}{\mathbf{p}(\mathbf{y})}\right) \\
& \text { if } \mathbf{p}(\mathbf{y}) \leq \mathbf{p}(\mathbf{x})\end{cases}  \tag{11}\\
\mu_{\mathbf{A}}(\mathbf{x}, \mathbf{y})=\frac{2|\mathbf{x} \cap \mathbf{y}|}{|\mathbf{x}|+|\mathbf{y}|} \tag{10}
\end{gather*}
$$

Let $|x|$ and $|y|$ be the frequency of concordances which contains the bigram pairs x and y , respectively. $|x \cap y|$ means how often two pairs x and $y$ co-occur in the same concordances under the distance constraint. (10) is relative entropy measure and (11) is dice coefficient. This measures are concerned with a lexical relation for cohesive degrees.

To get equivalence sets, it is very important to identify properties of the relation $R$ we defined. A relation which is reflexive, symmetric and transitive is called as an equivalence relation or similarity relation. In our case, the fuzzy cohesive relation, $R$ is certainly reflexive and symmetric. If $R(x, z) \geq \max _{y \in Y}$ $\min [R(x, y), R(y, z)]$ is satisfied for all $(x, z) \in$ $X^{2}$, then $R$ is transitive. Generally, transitive closure is used for checking transitivity. The transitive closure of a relation is defined as the smallest fuzzy relation which is transitive and has the fewest possible members with containing the relation itself.

Given a relation $S(X, X)$, its max-min transitive closure $S_{T}(X, X)$ can be calculated by the following algorithm consisted of three steps:

1. $S^{\prime}=S \cup(S \circ S), \circ$ is a max-min composition operator.
2. If $S^{\prime} \neq S$, make $S=S^{\prime}$ and go to Step 1 .
3. Stop: $S^{\prime}=S_{T}$.

If above algorithm terminates after the first iteration when applied to $R, R$ satisfies transitivity. To verify its transitivity, above alogrithm were employed. As a result, $R$ did not satisfy transitivity. It means that an element of $X$ could be-
long to multiple classes by $R$. This proves that the relation $R$ is valid to explain collocations.

A fuzzy binary relation $R(X, X)$ which is reHexive and symmetric is called as a fuzzy compatibility relation and is usually referred to as a. (puasi-equivalence relation. When $R$ is a fuzzy rompatibility relation, compatibility classes are defined in terms of a specified membership degree $\varepsilon$. An $\alpha$-compatibility class is a subset $A$ of $X$. such that $R(x, y) \geq \alpha$ for all $x, y \in A$ and the family consisting of the compatibility classes is called as an $\alpha$-cover of $X$ to $R$ in terms of a specific membership degree $\alpha$. An $\alpha$-cover forms partitions of $X$ and an element of $X$ could belong to multiple $\alpha$-compatibility classes. Here, we accepted $\alpha$-covers at $0.20 \alpha$-level in dice and (0.30 in relative entropy.

One might argue why we did not directly apply all bigrams to this stage with skipping the previous stage. We hope to deal with the comparision in a later paper.

## 5 Evaluation

We performed experiments for evaluation on 328,859 sentences( 8.5 million-morphemes) from Yonsei balanced copora. 250 morphemes were selected for a test, such that frequency $\geq 150$. The morphemes have 8,064 pairs and 773 were extracted as meaningful bigrams. In the second stage, 3,490 disjoint $\alpha$-compatibility classes corresponding to lexically cohesive clusters were generated. 698 longest n -gram collocations out of the $\alpha$-compatibility classes were extracted by eliminating the fragments that can be subsumed in longer classes.

The precision of extracted meaningful bigram was $86.3 \%$ and $92 \%$ in the case of n-gram collocations. We could take either $\alpha$-covers and the longest n-grams as n-gram collocations according to applications.

Since unfortunately, there is no existing database of collocations for evaluation, it is not easy to compute precision values and recall valnes as well. We computed the precision values by hand. As a different approach to Korean collocations, (Lee et al., 1996) extracted interrupted bigrams using several filtering conditions and at least the $90 \%$ of the results were adjacent bigrams of length 1 . By this comparison, we may conclude that our approach is more flexible to deal with Korean word order.

Figure $3{ }^{9}$ displays the changes of rank according to measures we considered. It shows that in contrast to other models, the properties have been effective in retrieving collocations which contain pairs of morphemes with relatively low frequency. Since the ranks of bigrams in four measures came up with our expectation, if we could make more adequate evaluation function, the precision would be improved.

Table 4 shows some obtained meaningful bigrams of '아니(not)'. There are a great deal of expressions relating negative sentences in Korean. The components of them occurs separated in various ways. When evaluating meaningful bigrams, the coefficients for the evaluation function are as follows: $\phi_{r} \approx 0.432, \phi_{c} \approx 0.490$, $\phi_{c r} \approx 0.371$ in the case of '아니(not)'. This means that the influence of three other measures is 1.284 times more than that of frequency measure in 'JP' POS relation.

We will illustrate all steps with a word, '신'(wear). The results of the first stage, meaningful bigrams of '신' (wear) ${ }^{10}$ are shown in Figure 4 . In the second stage, we calculated membership grades of inputs using dice measure and relative entropy measure. As Figure 4 shows, dice measure looks unsatisfactory in such cases as the pair '(를(object case), 많이(much))'. Although the common frequency, 3 is a relatively high in the aspect of the word with lower frequency, '많이'(much), the value of dice is low. Thus, we also tested relative entropy based on the probability of low frequency. Two measures produce similar results if all values in the level set of $R$ is considered instead of a specific value of $\alpha$, but entropy measure produces more good results.

Figure 4 and 5 show all $\alpha$-compatibility classes and the longest n-gram collocations of '신'(wear). Through our method, various kind of collocations were extracted. In Figure 4, the order of components of a $\alpha$ is by concordances.

## 6 Conclusion

In this paper, we implemented measures which reflect the four properties of collocation respec-

[^4]

Figure 3: Top 15 bigrams of '아나'(not) by our algorithm


Figure 4: Meaningful bigrams and all $\alpha$-compatibility classes of '신'(wear)
tively and the evaluation function which appropriately combines the measures. Our approach was primarily focused on the subtle relationships between word positions and collocations in a free order language.

We extracted meaningful bigrams using an evaluation function and extended them into
n-grams by producing $\alpha$-compatibility classes. The usefulness of our algorithm were illustrated by examples and tables.

This method covered various range of collocations, which the extracted collocation patterns were case frames, multiple functional words, selectional restrictions, semantic phrases, com-

| (Adverb,'아니') | (Postposition,'아니') | (Noun, '아니') |
| :--- | :--- | :--- |
| (비단, 아니)(not only) | (뿐, 아니)(too) | (때문, 아니)(not because) |
| (단지, 아니)(not only) | (만이, 아니)(not also) | (것, 아니)(not $\sim$ that) |
| (단순히, 아니)(simply $\sim$ not) | (가, 아니)(not) | (문제, 아니)(not important) |
| (그러나, 아니)(but $\sim$ not) | (도, 아니)(not) | (뚯, 아니)(not intend to) |
| (결코, 아니)(never) | (는, 아니)(not) | (예기, 아니)(not $\sim$ that) |
| (반드시, 아니)(not necessarily) | (만, 아니)(not also) | (이유, 아니)(not the reason) |

Table 4: Examples of bigrams for negation expressions


Figure 5: the longest n-gram collocations of '신'(wear)
pound nouns, and idioms and it could be applicable to other free order languages.

With the development of recognition of phrases, the input format and related distance between morphemes, the algorithm can be used effectively. Also linguistic contents for statistical constraints should be reflected in the system.

We have plans to check how this algorithm will work in English and to align bilingual collocations for machine translation.

## References

Benson, M., Benson, E., Ilson. R. 1986 The BBI Combinatory Dictionary of English: A Guide to Word Combinations. John Benjamins, Amsterdam and Philadelphia.
Breidt, E. 1993 Extraction of V-N collocations from text corpora: A feasibility study for German. In the 1st ACL-Workshop on Very Large Corpora.
Choueka, Y., Klein, T., and Neuwitz, E. 1983. Automatic retrieval of frequent idiomatic and collocational expressions in a large corpus. In Journal for Literary and Linguistic Computing, 4:34-38.
Church, K., and Hanks, P. 1989. Word association norms, mutual information, and
lexicography. In Computational Linguistics, 16(1):22-29.
Cowie, A. P. 1981. The treatment of collocations and idioms in learner's dictionaries. In Applied Linguistics, 2(3):223-235.
Cruse, D. P. 1986 Lexical Semantics. Cambridge University Press.
Haruno, M., Ikehara, S., and Yamazaki. T. 1996 Learning bilingual collocations by word-level sorting. In Proceedings of the 16 th COLING, 525-530.
Ikehara, S., Shirai, S., and Uchino, H. 1996 A Statistical Method for Extracting Uninterrupted and Interrupted Collocations. In Proceedings of the 16th COLING, 574-579.
Kjellmer, G. 1995 A mint of phrases: Corpus Linguistics, 111-127. Longman.
Klir, J. G., and Yuan, B. 1995 Fuzzy Sets And Fuzzy Logic: Theory And Applications. Prentice-Hall.
Martin, W., and Sterkenburg, V. P. 1983 Lexicography: principles and practice
Nagao, M., and Mori, S. 1994. A new method of $n$-gram statistics for large number of $n$ and automatic extraction of words and phrases from large text data of Japanese. In Proceedings of the 15th COLING, 611-615.
Ross, S. M. 1987 Introduction To Probability

And Statistics For Engineers And Scientists John Wiley \& Sons
Shimohata, S., Sugio, T., and Nagata, J. 1997. Retrieving collocations by co-occurrences and word order constraints. In the 35th Annual Meeting of ACL, 476-481.
Smadja, F. 1993. Retrieving collocations from text: Xtract. In Computational Linguistics, 19(1):143-177.
Smadja, F., MaKeown, K., and Hatzivassiloglou, V. 1996. Translating collocations for bilingual lexicons: A statistical approach. In Computational Linguistics, 22(1):1-38.
Lee, KongJoo, Kim, Jaehoon Kim, Kim, Gilchang. 1995. Extracting Collocations from Tagged Corpus in Korean. In Proceedings of Korean Information Sience Society, 22(2):623-626.


[^0]:    ${ }^{1}$ Bigrams and n-grams can be either adjacent morphemes or separated morphems by an arbitrary number of other words.
    ${ }^{2}$ In the case of an interrupted collocation, words can be separated by an arbitrary number of words, whereas

[^1]:    ${ }^{3}$ 'Noun','adJective','Verb','Postposition', 'aDverb', 'Ending', 'pre-ending', 'cOpular', 'Conjunction', 'Auxiliary verb', 'Suffix', 'etc.'.

[^2]:    ${ }^{4}$ We ported Smadja's Xtract tool into a Korean version.

[^3]:    ${ }^{5}$ Idioms have no ambiguous meaning but requires rigid patterns to preserve the idiomatic meaning.
    ${ }^{6}$ The replacement of some components by other words is more free than idioms.
    ${ }^{7}$ The combination of words is affected by selectional restrictions of predicate, noun, or adverb.
    ${ }^{8}$ It corresponds to multiple function word and appears on a adjacent word group.

[^4]:    ${ }^{9}$ The meanings of pairs are not described in detail because the pairs including function words are hard to translate into English.
    ${ }^{10}$ The word meaning corresponds to "put on(wear or take on)" in English, but it uses for shoes or socks.

