Recognizing Korean Unknown Proper Nouns by Using Automatically Extracted Lexical Clues

Bong-Rae Park, Young-Sook Hwang, Hae-Chang Rim
Natural Language Processing Lab,
Department of Computer Science and Engineering, Korea University,
Anam-Dong, Seoul 136, Republic of South Korea
pbr@nlp.korea.ac.kr, yshwang@nlp.korea.ac.kr, rim@nlp.korea.ac.kr

Apstract

This paper presents a method of extracting lexical clues automatically from a very large corpus and recognizing unknown proper nouns by using those lexical clues. This method collects proper noun candidates from the raw corpus and extracts the lexical clues among the adjacent known words of the proper noun candidates. And then, it recognizes unknown nouns and determines whether the identified unknown noun is a proper noun or not by using its adjacent lexical clues. Experimental result shows that the proposed method can extract 1,416 lexical clues from about ten million word size corpus and can recognize unknown proper nouns in the test corpus in 92% precision rate and 72% recall rate respectively.

1. Introduction

Many current application systems of natural language processing have been developed based on the assumption that all words within texts are registered in a machine-readable dictionary. But, this assumption is wrong because there are many unknown words in real texts (Park 1997) (Lee 1995) (Weischedel 1993).

In Korean, an unknown word can be a proper noun, an affix-derived word, or a foreign noun, etc. Each kind of unknown words has different problems in being recognized. Especially, recognition of an unknown proper noun has two critical problems. The first problem is that an unknown proper noun is difficult to recognize in a word level analysis because a proper noun is classified according to its meaning rather than its grammatical function. Moreover, the Korean proper nouns don't have any surface features unlike the other language; English proper nouns use an uppercase initial which is useful to recognize unknown proper nouns, but Korean proper nouns don't have such a property. Therefore, recognition of unknown proper nouns requires a kind of context analysis beyond a word level analysis. And the second problem is that many proper nouns temporarily appear on texts, so it is inappropriate to register them in a dictionary even if they are recognized. If we register temporarily used personal names or place names in a lexical dictionary as soon as they are recognized, then the dictionary becomes inefficiently large and causes many improper morphological analyses (Park 1995) (Atwell 1987). Accordingly, unknown proper nouns must be recognized in a real time without depending on a dictionary.

1.1. Existing Approach

Two existing methods are well known for dealing with Korean unknown proper nouns. The first method is to split a josa[1] from an eojeol[2] which fails to be morphologically analyzed and then to regard the head of the eojeol as an unknown proper noun. And the second method tries to recognize unknown proper nouns by using manually extracted lexical clues.

The first method is based on the assumption that all eojeols including unknown proper nouns fail to be morphologically analyzed and all eojeols which fail to be mophologically analyzed include unknown proper nouns. However, we observed that some eojeols which fail to be morphologically analyzed do not include unknown proper nouns but the other kinds of unknown words or orthographic errors. Moreover, about 10% of eojeols including unknown proper nouns can be improperly analyzed[3], and the last syllable of an unknown proper noun can often be mistaken for the first syllable of a josa (Park 1997). Therefore, this method suffers from some difficulties in recognizing unknown proper nouns.

And the second method recognizes an unknown proper noun by using their adjacent lexical clues. In this method, lexical clues are manually extracted by human experts in advance. Therefore, this method requires labor intensive work (Yang 1996). Recently, a semi-automatic method has been reported to extract more lexical clues by using some reliable lexical clues which are prepared by human experts (Strzalkowski 1996). In this paper, we present an automatic method of extracting lexical clues.

1.2. System Overview

Our method of recognizing unknown proper nouns consists of two stages as shown in Figure 1. The first stage is to extract lexical clues, and the second stage is to recognize unknown proper nouns by using those lexical clues.

[1] A josa is a tail combined with a noun head in Korean.

[2] An eojeol is a spacing unit in Korean like a word in English. An eojeol consists of one or more morphemes. It sometimes corresponds to a word or a phrase in English.

[3] In our test corpus, 9.8% of eojeols including unknown proper nouns are improperly analyzed.

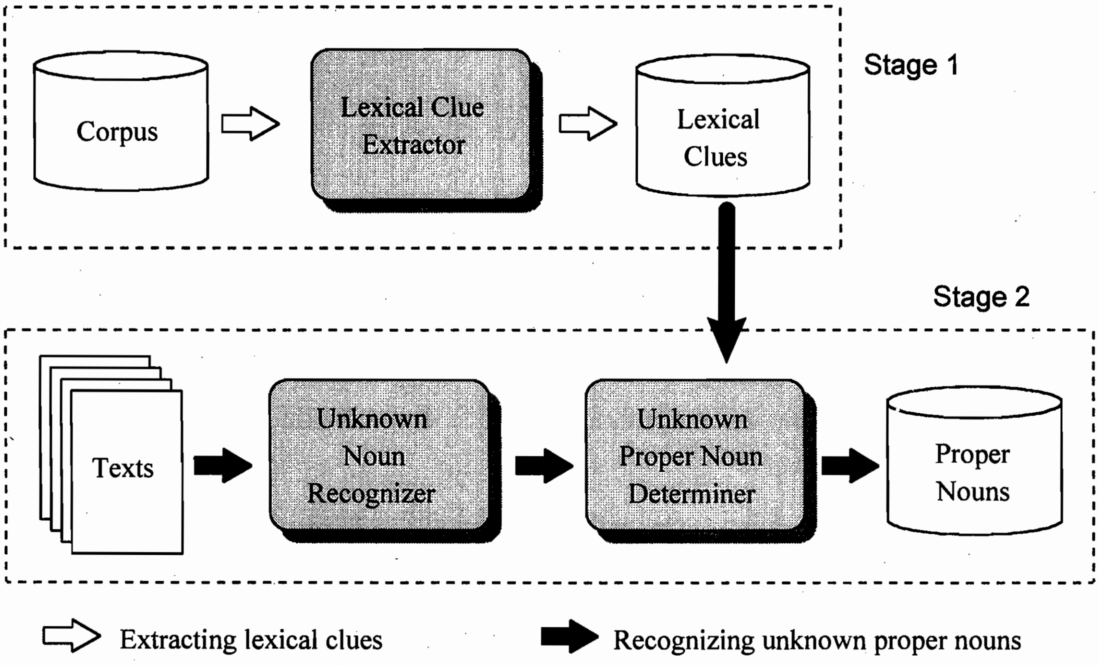

Figure 1. System configuration

The first stage is implemented by lexical clue extractor. The lexical clue extractor collects all eojeols which fail to be morphologically analyzed, and selects eojeols which include proper noun candidates. And then it extracts lexical clue candidates from the adjacent known words of the proper noun candidates, and determines whether each lexical clue candidate is a real lexical clue by estimating its degree of coupling with proper nouns. The degree of coupling with proper nouns is estimated by using two probabilities: the probability that the lexical clue candidate occurs in a set of the eojeols including the proper noun candidates and the probability that the lexical clue candidate occurs in an entire corpus. If the former probability is larger than the latter probability, the lexical clue candidate is determined to be a real lexical clue which can be used to recognize unknown proper nouns.

And the second stage is implemented by two processors: unknown noun recognizer and unknown proper noun determiner. The unknown noun recognizer is a kind of a preprocessor for recognizing unknown proper nouns and its function is to extract unknown words from unknown eojeols and identify unknown nouns among the extracted unknown words. And then, the proper noun determiner determines unknown proper nouns among the unknown nouns. In other words, this processor detects unknown nouns which occur together with one or more lexical clues and determines whether each unknown noun is an unknown proper noun by using its degree of coupling with the adjacent lexical clues.

2. Extracting Lexical Clues

Our method recognizes unknown proper nouns by using their lexical clues. Therefore, the extraction of qualified lexical clues has a good effect on precision and recall rates of recognizing unknown proper nouns. This section presents a method of extracting lexical clues automatically from a very large raw corpus. First, we collect eojeols which are expected to include unknown proper nouns. Most eojeols which include unknown proper nouns fail to be morphologically analyzed. So, we gather all eojeols which fail to be morphologically analyzed, and then filter out the eojeols including affix-derived words[4] or spacing errors (Park 1995) from them[5]. The eojeols including affix-derived words can be found by the analysis of one-syllable affix[6]. Also, the eojeols including spacing errors can be detected by using the existing method of detecting spacing errors. Accordingly, the remaining eojeols are expected to include unknown proper nouns.

In Korean, an eojeol can be splitted into a head and a tail, in which a head consists of one or more lexical morphemes and a tail consists of zero or more grammatical morphemes. A proper noun and a lexical clue are often combined to become a head because they are lexical morphemes. Therefore, we split heads from the above collected eojeols which are expected to include proper nouns and select only heads having both a proper noun and a lexical clue.

Figure 2 shows a detailed algorithm of collecting those heads and extracting lexical clues from those heads. In the steps 1 to 3, we collect eojeols which fail to be morphologically analyzed and filter out eojeols including affix-derived words and spacing errors from them, and so remaining eojeols are expected to include only proper nouns. And then, in the steps 4 to 8, we extract lexical clue candidates from the adjacent known words of the proper noun candidates and measure their occurrence probabilities that the lexical clue candidates occur near proper noun candidates. And in the steps 9 to 11, we measure the occurrence probabilities that the lexical clue candidates occur in the entire corpus. Finally, in the step 12, we compare two occurrence probabilities of each lexical clue candidate and extract the lexical clues which occur more often near proper noun candidates than the other words.

[4] In Korean, affixes are very diverse and difficult to distinguish from the other words.

[5] In Korean, most eojeols which fail to be morphologically analyzed include an unknown proper noun, an unknown affix-derived word, or a spacing error.

[6] Generally, one-syllable affix analysis has not been performed in Korean language processing systems because it causes the overgeneration problem.

```
// Collecting eojeols including proper nouns
1: Collect all eojeol E's that fail to be morphologically analyzed from corpus R.
2: Exclude eojeol E's with one syllable affix.
3: Exclude eojeol E's having any spacing error.
// Extracting lexical clue candidates and their occurrence probabilities
4: Extract head H's from the remaining eojeol E's.
5: Exclude head H's whose length is less than four syllables.
6: Construct a set X with head H's.
7: Extract lexical clue candidate c's from each head H in the set X.
8: Estimate a probability P(c|X) for each lexical clue candidate.
// Estimating occurrence prob.'s of the lexical clue candidates in a raw corpus
9: Extract all head S's from all eojeols within the corpus R.
10: Construct a set A with heads S's.
11: Estimate a probability P(c|A) for each lexical clue candidate c.
// Selecting lexical clues by comparing above two kinds of probabilities.
12: Select lexical clue candidates according to the following formula.
        $\frac{P(c|X)}{P(c|A)} > 1$
```

Figure 2. An algorithm of selecting lexical clues

3. A method of recognizing unknown nouns

In Korean, unknown words can be classified into nouns, verbs and adverbs according to their part-of-speeches. The number of unknown verbs and adverbs are small, but they can not be neglected, and a proper noun is a kind of a noun. Therefore, unknown nouns must be recognized from the unknown words before we try to recognize unknown proper nouns.

The existing methods of recognizing unknown nouns detect eojeols which fail to be morphologically analyzed and generate every possible unknown word candidates from the eojeols and then select optimal unknown word candidates by using stochastic and/or linguistic informations. However, the existing methods have three critical problems. The first problem is that those methods can't detect any unknown word candidates from improperly analyzed eojeols which include unknown words, and the second problem is that those methods often overgenerate unknown word candidates from eojeols which fail to be morphologically analyzed. And third problem is that those methods have difficulty in splitting an unknown noun and known words in the same eojeol.

Figure 3 shows the basic idea of the existing methods[7]. In this figure, eojeols 이순신의[*lee-sun-sin-eui*] and 원정가면[*won-jeong-ga-myeon*] have unknown words 이순신[*lee-sun-sin*] and 원정가[*won-jeong-ga*] respectively, but those unknown words are not detected because the eojeols are improperly analyzed[8]. And two or more unknown word candidates are overgenerated from eojeols 원정가서도[*won-jeong-ga-seo-do*] and 이순신장군만[*lee-sun-sin-jang-gun-man*] which fail to be morphologically analyzed. Moreover, the unknown noun 이순신[*lee-sun-sin*] is not exactly extracted from the eojeol 이순신장군만[{*lee-sun-sin-jang-gun-man*], but 이순신장군[*lee-sun-sin-jang-gun*] is extracted.

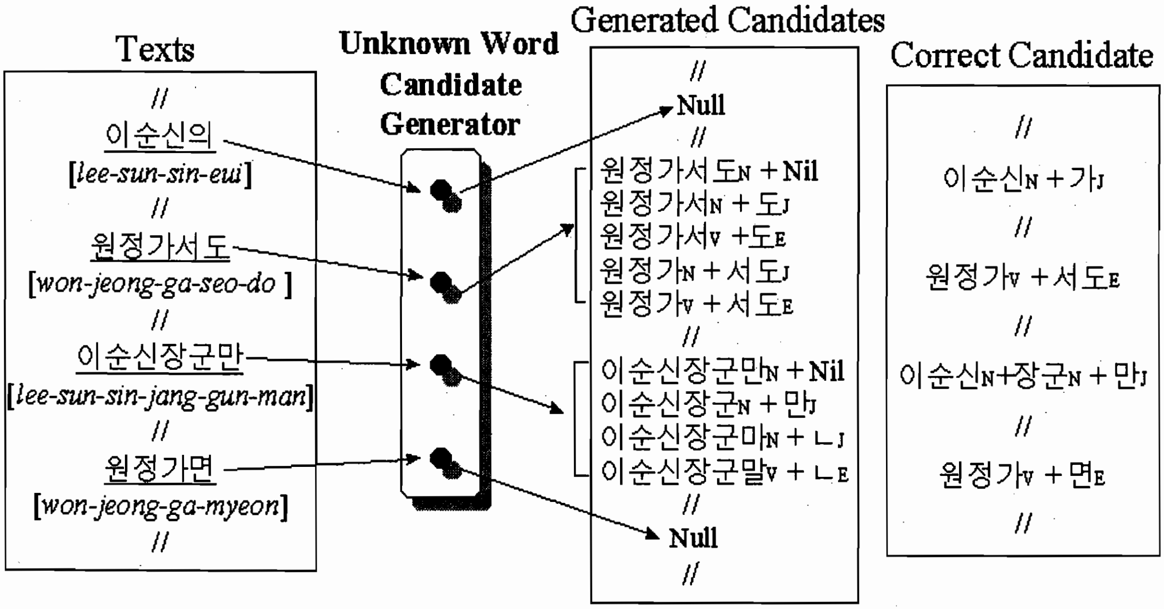

Figure 3. The existing unknown noun recognizing method

Our approach, named an example analysis method (Park 1997), is based on a comparative analysis of several example eojeols. We accept a candidate for an unknown word only if the candidate is consistently applied to its example eojeols.

[7] The tag 'N' stands for a noun and the tags 'J' and 'E' stand for a josa and an eomi respectively. In this case, an eomi is a tail combined with a verb head in Korean.

[8] The eojeol 이순신의[*lee-sun-sin-eui*] should be analyzed into 이순신$_N$+의$_J$, but it is improperly analyzed into 이순$_N$+신의$_N$, and the eojeol 원정가면[*won-jeong-ga-myeon*] should be analyzed into 원정가$_V$+면$_E$, but it is improperly analyzed into 원정$_N$+가면$_N$.

Figure 4 shows the basic idea of the example analysis method. The example analysis method comparatively analyzes the example eojeols 이순신의[*lee-sun-sin-eui*] and 이순신장군만[*lee-sun-sin-jang-gun-man*] and then recognizes the unknown noun 이순신 [*lee-sun-sin*] uniquely. Also, this method comparatively analyzes the example eojeols 원정가서도[*won-jeong-ga-seo-do*] and 원정가면[*won-jeong-ga-myeon*] and then recognizes the unknown verb 원정가[*won-jeong-ga*] uniquely.

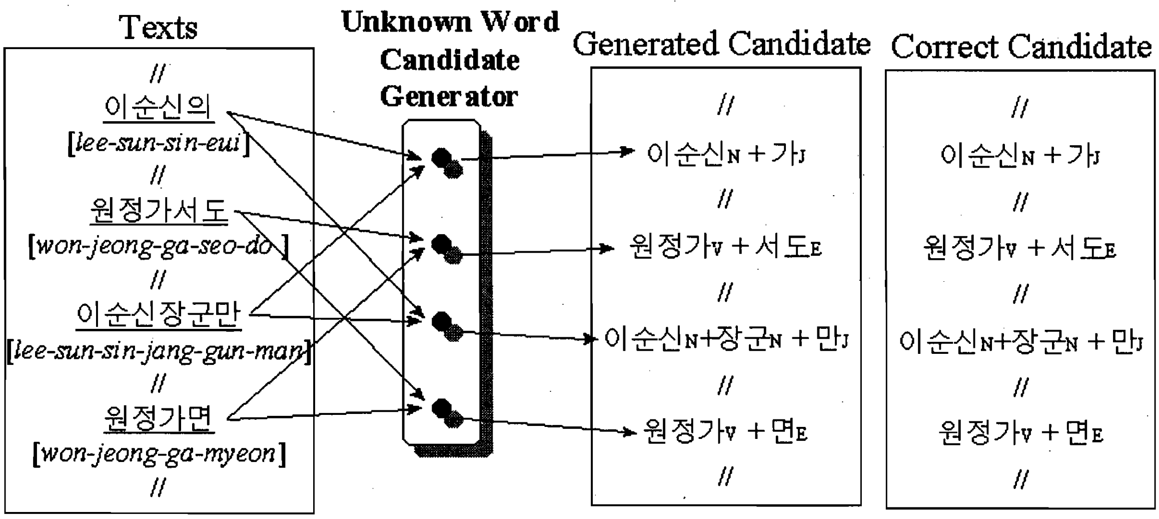

Figure 4. The example analysis method

Moreover, the example analysis method has a good effect especially on recognizing unknown proper nouns by using their lexical clues because it can effectively collect all distributed lexical clues of each unknown proper noun. Many proper nouns occur more than two places in their source text. Thus, two or more different lexical clues can appear together with an unknown proper noun. Our proper noun recognition method uses every distributed lexical clues to recognize unknown proper nouns.

4. Recognizing Unknown Proper Nouns

As mentioned in the section 2, the method of extracting lexical clues is to construct the set of heads including proper noun candidates(X) and the set of all heads(A), and then regard the known word(c) as a lexical clue for recognizing the proper noun when the occurrence probability of the known word(c) in the set(X), P(c|X), is larger than the occurrence probability of the known word(c) in the set(A), P(c|A). But, all the extracted lexical clues don't have equal clue powers because the P(c|X), P(c|A) and P(c|X)/P(c|A) can be different among the lexical clues. Therefore, it is necessary to estimate clue powers of each lexical clue and use these lexical clues differently according to their clue powers.

The clue power is estimated based on the following two assumptions. The first assumption is that a lexical clue(c) with high probability P(c|X) guarantees the recognition of many proper nouns. And the second assumption is that a lexical clue(c) with high value P(c|X)/P(c|A) guarantees the exact recognition of proper nouns. Therefore, the clue power(CPower, henceforth) is defined as follows:

$$\text{CPower(c)} = P(c|X) * \frac{P(c|X)}{P(c|A)} = \frac{P(c|X)^2}{P(c|A)} \qquad (1)$$

Equation (1) deals with only one lexical clue, but an unknown noun can have two or more distributed lexical clues. Therefore, we need to estimate the combined clue power of two or more lexical clues. So, we extend Equation (1) to Equation (2) assuming that the combined clue power of two or more lexical clues is equal to the summation of the clue powers of the componet lexical clues[9].

$$\text{CPower}(c_1, c_2, \ldots, c_n) = \sum_{i=1}^{n} CPower(c_i) \qquad (2)$$

For example, if lexical clues *meseum* and *curator* occur near a proper noun candidate, the combined clue power of these lexical clues for this proper noun candidate is as follows:

$$\text{CPower(museum, curator) = CPower(museum) + CPower(curator)}$$

By using two or more lexical clues together, even lexical clues with low clue powers[10] can be used in recognizing unknown proper nouns.

After we estimate the clue power of lexical clues of an unknown noun, we decide whether the unknown noun is a proper noun or not by comparing the clue power of its lexical clues with the predetermined threshold value as shown below.

$$\text{CPower}(c_1, c_2, \ldots, c_n) > T$$

[9] In equation (2), c_i is the i-th lexical clue of the unknown noun.

[10] There are two cases that a good lexical clue has a low clue power. The first case is that a data sparseness problem causes the factor P(c|X) of the clue power to be low. And the second case is that a lexical clue word has multiple meanings and only one meaning of them implies a clue. This case causes the factor P(c|X)/P(c|A) of the clue power to be low. However, these lexical clues are also used to recognize unknown proper nouns if two or more lexical clues are used together and their combined clue power is above the threshold.

Determination of a threshold value is an important factor to recognize unknown proper nouns. In this paper, we determine the threshold value based on the recall rate. In an ideal case, all lexical clues extracted from the set of heads including proper noun candidates(X) guarantees 100% recall rate. This percentage corresponds to the summation of occurrence probabilities of all the lexical clues over the set(X). That is, we can say each lexical clue affects the increment of the recall rate(R) by its occurrence probability P(c|X). Therefore, we determine the appropriate threshold(T) according to the required recall rate(R) as follows:[11]

$$T = \mathrm{CPower}(C_k) \quad \text{where } \sum_{i=1}^{k} P(c_i|X) > R \text{ and } \sum_{i=1}^{k-1} P(c_i|X) \leq R \qquad (k \leq N)$$

$$P(c_i|X) \geq P(c_j|X) \quad (i > j)$$

According to this formula, the threshold is decided to be CPower(c_k) when lexical clues(c_i) are sorted in the descending order according to their ccurrence probabilities, P(c_i|X), and the summation of P(c_1|X) to P(c_k|X) is above the recall rate, but the summation of P(c_1|X) to P(c_{k-1}|X) is not above the recall rate.

5. Experiment

5.1. Extraction of lexical clues

We extracted 274,682 unique eojeols which fail to be morphologically analyzed from 10 million eojeol size corpus. From those eojeols, we excluded eojeols having affix-derived words and spacing errors, and constructed the set(X) with the unique heads of the remaining eojeols. And then, we selected 5,486 lexical clue candidates from the set(X) and estimated their occurrence probabilities in the set(X). Also, we constructed the set(A) with 563,057 unique heads of the entire corpus and estimated the occurrence probabilities of the lexical clue candidates in the set(A). And we acquired 1,416 lexical clues by comparing the occurrence probabilities of the lexical clue candidates in those sets(X and A). The CPowers of 503 lexical clues among them are above the threshold with 80% recall rate[12]. The table 1 shows the example of lexical clues extracted with high clue powers.

[11] N is the number of all lexical clues.

[12] This means that unknown nouns with such a lexical clue in their neighbor are recognized as unknown proper nouns. The other lexical clues are used to recognize unknown proper nouns only if two or more lexical clues are used together and their combined clue power is above the threshold.

Table 1. The example of lexical clues

Lexical clue	Meaning	Lexical clue	Meaning
의원[*eu-won*]	member	선생[*seon-saeng*]	teacher
장관[*jang-gwan*]	minister	출신[chul-sin]	affiliation
대표[*dae-pyo*]	delegate	그룹[*geu-rup*]	business group
변호사[*byeon-ho-sa*]	lawyer	아파트[*a-pa-t*]	apartment
총리[*chong-lea*]	premier	사장[*sa-jang*]	president of company
부장[*bu-jang*]	manager	지역[*ji-yeok*]	district
박사[*bak-sa*]	doctor	대학[*dae-hak*]	university
병원[*byeon-won*]	hospital	정권[*jeong-gwon*]	regime
교수[*gyo-su*]	professor	대변인[*dae-byeon-in*]	spokesman
대통령[*dae-tong-lyeong*]	president	위원[*wi-won*]	committee
총장[*chong-jang*]	president of university	은행[*eun-haeng*]	bank
검사[*geom-sa*]	prosecutor	백화점[*baek-hwa-jeom*]	department store

5.2. Recognition of proper nouns

To verify the proposed method, we used 120,000 word size test corpus collected from newspapers and novels. Figure 5 shows the distribution of unknown proper nouns within the test corpus and Table 2 shows the comparison between the josa splitting method and the proposed method[13].

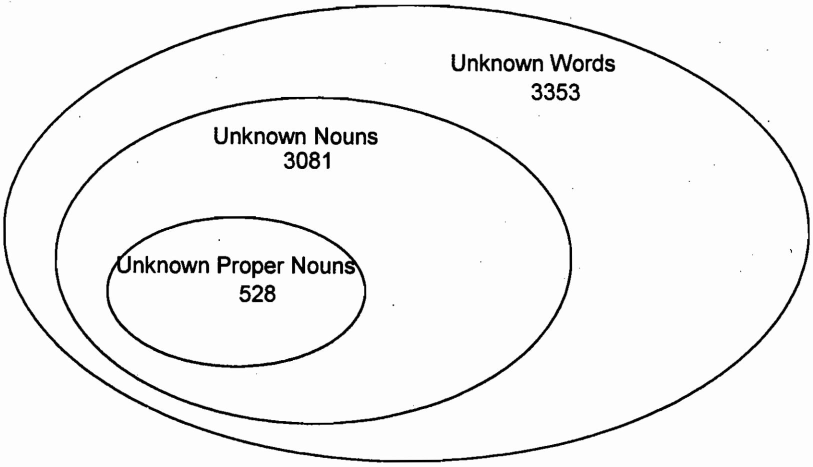

Figure 5. Distribution of unknown proper nouns

[13] The threshold applied to the proposed method of recognizing unknown proper nouns is determined when the expected recall rate is 80%.

Table 2. The comparison between the josa splitting method and the proposed method

Item	Josa splitting method		Proposed method	
	recall	precision	recall	precision
Unknown nouns	76.2	87.0	86.3	94.7
Unknown proper nouns	90.2	20.3	71.7	92.4

According to Table 2, the josa splitting method is superior to the proposed method in terms of the recall rate, but the proposed method is much superior to the josa splitting method in terms of precision. And, in Table 2, the recall rate of the josa splitting method may be expected to be almost 100% because this method regards all noun candidates as proper nouns, but it turns out to be only 90.2% because the josa splitting method can't recognize unknown proper nouns from improperly analyzed eojeols and such eojeols are 9.8% of all eojeols including unknown proper nouns. And the precision rate of the josa splitting method is very low because only about 20% of all unknown nouns were proper nouns at least in this test corpus. On the other hand, the precision rate of the proposed method is very high. And moreover, the proposed method can recognize unknown proper nouns from improperly analyzed eojeols and split an unknown proper noun and known words in the same eojeol.

6. Conclusion and Future Work

In this paper, we have presented a method of recognizing unknown proper nouns by using automatically extracted lexical clues. Our method consists of two stages. The first stage is to extract stochastically lexical clues which prefer to occur with proper nouns. And the second stage is to recognize unknown nouns having one or more lexical clues in their neighborhood and determines whether the unknown nouns are proper nouns or not by applying the given threshold to the clue power of those lexical clues. Experimental result shows that our method extracts 1,416 lexical clues from about ten million word size raw corpus, and recognizes unknown proper nouns in 92% precision rate and 72% recall rate respectively.

In the future work, we will cluster the selected lexical clues by Kohonen's SOFM(Self-Organizing Feature Map) (Pandya 1996) to recognize unknown proper nouns according to their categories. And, we will try to extend the scope of extracting lexical clues to the adjacent eojeols which are located near the eojeols including unknown proper noun candidates.

Reference

Atwell, Eric, Stephen Elliott, "Dealing with ill-formed Englisth text," *The computational Analysis of English: a corpus-based approach*, Longman, 1987, pp.120-138.

Lee, Sang-Ho, et al, "A Korean part-of-speech tagging system with handling unknown words," *Proc. of 1995 International Conference on Computer Processing of Oriental Languages*, Nov. pp.23-25, Honolulu.

Mikheev, Andrei, "Unsupervised Learning of Word-Category Guessing Rules," *Proc. of the 34th ACL*, 1996, pp.327-334.

Pandya, Abhijit S. and Robert B.Macy, *Pattern Recognition with Neural Networks in C++*, CRC Press, 1996.

Park, Bong-Rae, Hae-Chang Rim, "A Korean Corpus Refining System based on Automatic Analysis of Corpus," *Proc. of Natural Language Processing Pacific Rim Symposium*, 1995, pp.89-94.

Park, Bong-Rae, Young-Sook Hwang, Hae-Chang Rim, "Recognizing Korean Unknown Words by Comparatively Analyzing Example words," Proc. of 1997 *International Conference on Computer Processing of Oriental languages*, pp.127-132, Hong Kong.

Strzalkowski, Tomek and Jin Wang, “A Self-Learng Univeral Concept Spotter,” In *Proceedings of COLING-96*, 1996, pp.931-936.

Weischedel, Ralph, Marie Meteer, Richard Schartz, Lance Ramshaw, "Coping with Ambiguity and Unknown Words through Probabilistic Models," *Computational Linguistics*, Vol.19, 1993, pp.360-382.

Yang, Jang-Mo, Min-Jung Kim, Hyuk-Chul Kwon, "Extraction Method of the Unknown-Words with Linguistic Knowledge in Korean," *Proc. of Spring Conference of Korean Information Science Society*, 1996, pp.925-928. (in Korean)