A Syllable Based Word Recognition Model for Korean Noun Extraction

Do-Gil Lee and **Hae-Chang Rim**
Dept. of Computer Science & Engineering
Korea University
1, 5-ka, Anam-dong, Seongbuk-ku
Seoul 136-701, Korea
{dglee, rim}@nlp.korea.ac.kr

Heui-Seok Lim
Dept. of Information & Communications
Chonan University
115 AnSeo-dong
CheonAn 330-704, Korea
limhs@infocom.chonan.ac.kr

Abstract

Noun extraction is very important for many NLP applications such as information retrieval, automatic text classification, and information extraction. Most of the previous Korean noun extraction systems use a morphological analyzer or a Part-of-Speech (POS) tagger. Therefore, they require much of the linguistic knowledge such as morpheme dictionaries and rules (e.g. morphosyntactic rules and morphological rules).

This paper proposes a new noun extraction method that uses the syllable based word recognition model. It finds the most probable syllable-tag sequence of the input sentence by using automatically acquired statistical information from the POS tagged corpus and extracts nouns by detecting word boundaries. Furthermore, it does not require any labor for constructing and maintaining linguistic knowledge. We have performed various experiments with a wide range of variables influencing the performance. The experimental results show that without morphological analysis or POS tagging, the proposed method achieves comparable performance with the previous methods.

1 Introduction

Noun extraction is a process to find every noun in a document (Lee et al., 2001). In Korean, Nouns are used as the most important terms (features) that express the document in NLP applications such as information retrieval, document categorization, text summarization, information extraction, and etc.

Korean is a highly agglutinative language and nouns are included in Eojeols. An Eojeol is a surface level form consisting of more than one combined morpheme. Therefore, morphological analysis or POS tagging is required to extract Korean nouns.

The previous Korean noun extraction methods are classified into two categories: morphological analysis based method (Kim and Seo, 1999; Lee et al., 1999a; An, 1999) and POS tagging based method (Shim et al., 1999; Kwon et al., 1999). The morphological analysis based method tries to generate all possible interpretations for a given Eojeol by implementing a morphological analyzer or a simpler method using lexical dictionaries. It may overgenerate or extract inaccurate nouns due to lexical ambiguity and shows a low precision rate. Although several studies have been proposed to reduce the over-generated results of the morphological analysis by using exclusive information (Lim et al., 1995; Lee et al., 2001), they cannot completely resolve the ambiguity.

The POS tagging based method chooses the most probable analysis among the results produced by the morphological analyzer. Due to the resolution of the ambiguities, it can obtain relatively accurate results. But it also suffers from errors not only produced by a POS tagger but also triggered by the preceding morphological analyzer.

Furthermore, both methods have serious deficien-

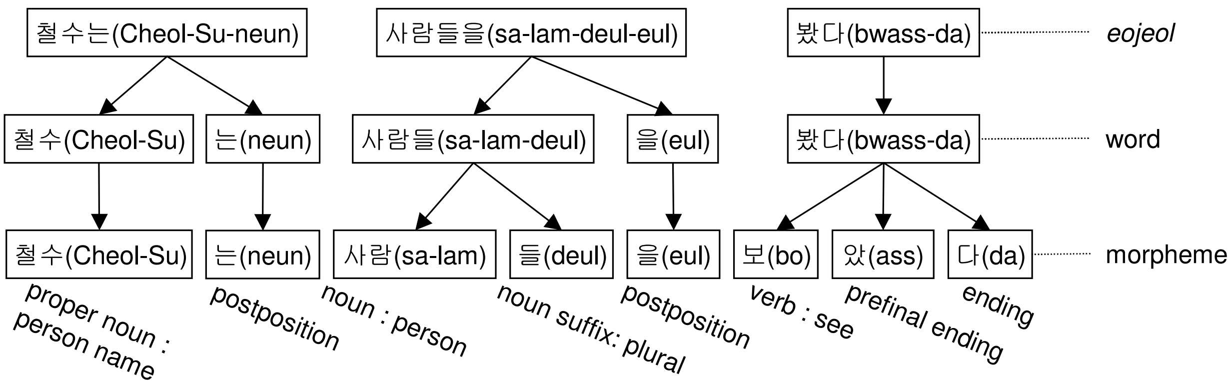

Figure 1: Constitution of the sentence "철수는 사람들을 봤다(Cheol-Su saw the persons)"

cies in that they require considerable manual labor to construct and maintain linguistic knowledge and suffer from the unknown word problem. If a morphological analyzer fails to recognize an unknown noun in an unknown Eojeol, the POS tagger would never extract the unknown noun. Although the morphological analyzer properly recognizes the unknown noun, it would not be extracted due to the sparse data problem.

This paper proposes a new noun extraction method that uses a syllable based word recognition model. The proposed method does not require labor for constructing and maintaining linguistic knowledge and it can also alleviate the unknown word problem or the sparse data problem. It finds the most probable syllable-tag sequence of the input sentence by using statistical information and extracts nouns by detecting the word boundaries. The statistical information is automatically acquired from a POS annotated corpus and the word boundary can be detected by using an additional tag to represent the boundary of a word.

This paper is organized as follows. In Section 2, the notion of word is defined. Section 3 presents the syllable based word recognition model. Section 4 describes the method of constructing the training data from existing POS tagged corpora. Section 5 discusses experimental results. Finally, Section 6 concludes the paper.

2 A new definition of word

Korean spacing unit is an Eojeol, which is delimited by whitespace, as with *word* in English. In Korean, an Eojeol is made up of one or more words, and a word is made up of one or more morphemes. Figure 1 represents the relationships among morphemes, words, and Eojeols with an example sentence. Syllables are delimited by a hyphen in the figure.

All of the previous noun extraction methods regard a morpheme as a processing unit. In order to extract nouns, nouns in a given Eojeol should be segmented. To do this, the morphological analysis has been used, but it requires complicated processes because of the surface forms caused by various morphological phenomena such as irregular conjugation of verbs, contraction, and elision. Most of the morphological phenomena occur at the inside of a morpheme or the boundaries between morphemes, not a word. We have also observed that a noun belongs to a morpheme as well as a word. Thus, we do not have to do morphological analysis in the noun extraction point of view.

In Korean linguistics, a word is defined as a morpheme or a sequence of morphemes that can be used independently. Even though a postposition is not used independently, it is regarded as a word because it is easily segmented from the preceding word. This definition is rather vague for computational processing. If we follow the definition of the word in linguistics, it would be difficult to analyze a word like the morphological analysis. For this reason, we define a different notion of a word.

According to our definition of a word, each uninflected morpheme or a sequence of successive inflected morphemes is regarded as an individual word. [1] By virtue of the new definition of a word, we need not consider mismatches between the surface level form and the lexical level one in recognizing words.

The example sentence “철수는 사람들을 봤다(Cheol-Su saw the persons)” represented in Figure 1 includes six words such as “철수(Cheol-Su)”, “는(neun)”, “사람(sa-lam)”, “들(deul)”, “을(eul)”, and “봤다(bwass-da)”. Unlike the Korean linguistics, a noun suffix such as “님(nim)”, “들(deul)”, or “적(jeog)” is also regarded as a word because it is an uninflected morpheme.

3 Syllable based word recognition model

A Korean syllable consists of an obligatory onset (initial-grapheme, consonant), an obligatory peak (nuclear grapheme, vowel), and an optional coda (final-grapheme, consonant). In theory, the number of syllables that can be used in Korean is the same as the number of every combination of the graphemes. [2] Fortunately, only a fixed number of syllables is frequently used in practice. [3] The amount of information that a Korean syllable has is larger than that of an alphabet in English. In addition, there are particular characteristics in Korean syllables. The fact that words do not start with certain syllables is one of such examples. Several attempts have been made to use characteristics of Korean syllables. Kang (1995) used syllable information to reduce the over-generated results in analyzing conjugated forms of verbs. Syllable statistics have been also used for automatic word spacing (Shim, 1996; Kang and Woo, 2001; Lee et al., 2002).

The syllable based word recognition model is represented as a function Γ like the following equations. It is to find the most probable syllable-tag sequence $t_{1,n} = t_1, t_2, ..., t_n$, for a given sentence S consisting of a sequence of n syllables $c_{1,n} = c_1, c_2, ..., c_n$.

$$\Gamma(c_{1,n}) \overset{def}{=} \underset{t_{1,n}}{\operatorname{argmax}} P(t_{1,n} \mid c_{1,n}) \quad (1)$$

$$\approx \underset{t_{1,n}}{\operatorname{argmax}} \prod_{i=1}^{n} P(t_i \mid t_{i-1}) P(c_i \mid t_i) \quad (2)$$

Two Markov assumptions are applied in Equation 2. One is that the probability of a current syllable tag t_i conditionally depends on only the previous syllable tag. The other is that the probability of a current syllable s_i conditionally depends on the current tag. In order to reflect word spacing information in Equation 2, which is very useful in Korean POS tagging, Equation 2 is changed to Equation 3 which can consider the word spacing information by calculating the transition probabilities like the equation used in Kim et al. (1998).

$$\Gamma(c_{1,n}) = \underset{t_{1,n}}{\operatorname{argmax}} \prod_{i=1}^{n} P(t_i \mid t_{i-1}, k) P(c_i \mid t_i) \quad (3)$$

In the equation, k becomes zero if the transition occurs in the inside of an Eojeol; otherwise k is one.

Word boundaries can be detected by an additional tag. This method has been used in some tasks such as text chunking and named entity recognition to represent a boundary of an element (e.g. individual phrase or named entity). There are several possible representation schemes to do this. The simplest one is the BIO representation scheme (Ramshaw and Marcus, 1995), where a “B” denotes the first item of an element and an “I” any non-initial item, and a syllable with tag “O” is not a part of any element. Because every syllable corresponds to one syllable tag, “O” is not used in our task. The representation schemes used in this paper are described in detail in Section 4.

The probabilities in Equation 3 are estimated by the maximum likelihood estimator (MLE) using relative frequencies in the training data. [4]

The most probable sequence of syllable tags in a sentence (a sequence of syllables) can be efficiently computed by using the Viterbi algorithm.

[1] Korean morphemes can be classified into two types: uninflected morphemes having fixed word forms (such as noun, unconjugated adjective, postposition, adverb, interjection, etc.) and inflected morphemes having conjugated word forms (such as a morpheme with declined or conjugated endings, predicative postposition, etc.)

[2] $11,172 (= 19 \times 21 \times 28)$ of pure Korean syllables are possible

[3] Actually, $2,457$ of syllables are used in the training data, including Korean characters and non-Korean characters (e.g. alphabets, digits, Chinese characters, symbols).

[4] Since the MLE suffers from zero probability, to avoid zero probability, we just assign a very low value such as 1.0×10^{-100} for an unseen event in the training data.

Table 1: Examples of syllable tagging by BI, BIS, IE, and IES representation schemes

surface level (syllable)	lexical level (morpheme/POS tag)	BI	BIS	IE	IES
약(yak)	약속(yak-sok)/nc	B-nc	B-nc	I-nc	I-nc
속(sok)		I-nc	I-nc	E-nc	E-nc
장(jang)	장소(jang-so)/nc	B-nc	B-nc	I-nc	I-nc
소(so)		I-nc	I-nc	E-nc	E-nc
인(in)	이(i)/co+ ㄴ (n)/etm	B-co_etm	S-co_etm	E-co_etm	S-co_etm
신(Sin)	신라호텔(Sin-la-ho-tel)/nc	B-nc	B-nc	I-nc	I-nc
라(la)		I-nc	I-nc	I-nc	I-nc
호(ho)		I-nc	I-nc	I-nc	I-nc
텔(tel)		I-nc	I-nc	E-nc	E-nc
커(keo)	커피숍(keo-pi-syob)/nc	B-nc	B-nc	I-nc	I-nc
피(pi)		I-nc	I-nc	I-nc	I-nc
숍(syob)		I-nc	I-nc	E-nc	E-nc
에(e)	에(e)/jc	B-jc	S-jc	E-jc	S-jc
재(Jai)	재옥(Jai-Ok)/nc	B-nc	B-nc	I-nc	I-nc
옥(Ok)		I-nc	I-nc	E-nc	E-nc
이(i)	이(i)/jc	B-jc	S-jc	E-jc	S-jc
먼(meon)	먼저(meon-jeo)/mag	B-mag	B-mag	I-mag	I-mag
저(jeo)		I-mag	I-mag	E-mag	E-mag
와(wa)	오(o)/pv+아(a)/ec	B-pv_ec	S-pv_ec	E-pv_ec	S-pv_ec
기(gi)	기다리(gi-da-li)/pv+고(go)/ec	B-pv_ec	B-pv_ec	I-pv_ec	I-pv_ec
다(da)		I-pv_ec	I-pv_ec	I-pv_ec	I-pv_ec
리(li)		I-pv_ec	I-pv_ec	I-pv_ec	I-pv_ec
고(go)		I-pv_ec	I-pv_ec	E-pv_ec	E-pv_ec
있(iss)	있(iss)/px+었(eoss)/ep+다(da)/ef	B-px_ef	B-px_ef	I-px_ef	I-px_ef
었(eoss)		I-px_ef	I-px_ef	I-px_ef	I-px_ef
다(da)		I-px_ef	I-px_ef	E-px_ef	E-px_ef
.	./s	B-s	S-s	E-s	S-s

Given a sequence of syllables and syllable tags, it is straightforward to obtain the corresponding sequence of words and word tags. Among the words recognized through this process, we can extract nouns by just selecting words tagged as nouns. [5]

4 Constructing training data

Our model is a supervised learning approach, so it requires a training data. Because the existing Korean POS tagged corpora are annotated by a morpheme level, we cannot use them as a training data without converting the data suitable for the word recognition model. The corpus can be modified through the following steps:

Step 1 For a given Eojeol, segment word boundaries and assign word tags to each word.

Step 2 For each separated word, assign the word tag to each syllable in the word according to one of the representations.

[5]For the purpose of noun extraction, we only select common nouns here (tagged as "nc" or "NC") among other kinds of nouns.

In step 1, word boundaries are identified by using the information of an uninflected morpheme and a sequence of successive inflected morphemes. An uninflected morpheme becomes one word and its tag is assigned to the morpheme's tag. Successive inflected morphemes form a word and the combined form of the first and the last morpheme's tag represents its tag. For example, the morpheme-unit POS tagged form of the Eojeol "갔었다(gass-eoss-da)" is "가(ga)/pv+았(ass)/ep+었(eoss)/ep+다(da)/ef", and all of them are inflected morphemes. Hence, the Eojeol "갔었다(gass-eoss-da)" becomes one word and its tag is represented as "pv_ef" by using the first morpheme's tag ("pv") and the last one's ("ef").

In step 2, a syllable tag is assigned to each of syllables forming a word. The syllable tag should express not only POS tag but also the boundary of the word. In order to detect the word boundaries, we use the following four representation schemes:

BI representation scheme Assign "B" tag to the first syllable of a word, and "I" tag to the others.

BIS representation scheme Assign "S" tag to a syllable which forms a word, and other tags ("B" and "I") are the same as "BI" representation scheme.

IE representation scheme Assign "E" tag to the last syllable of a word, and "I" tag to the others.

IES representation scheme Assign "S" tag to a syllable which forms a word, and other tags ("I" and "E") are the same as "IE" representation scheme.

Table 1 shows an example of assigning word tag by syllable unit to the morpheme unit POS tagged corpus.

Table 2: Description of Tagset 2 and Tagset 3

Tag Description	Tagset 2	Tagset 3
symbol	s	S
foreign word	f	F
common noun	nc	NC
bound noun	nb	NB
pronoun	np	NP
numeral	nn	NN
verb	pv	V
adjective	pa	A
auxiliary predicate	px	VX
copula	co	CO
general adverb	mag	MA
conjunctive adverb	maj	
adnoun	mm	MM
interjection	ii	IC
prefix	xp	XPN
noun-derivational suffix	xsn	XSN
verb-derivational suffix	xsv	XSV
adjective-derivational suffix	xsm	
case particle	jc	J
auxilary particle	jx	
conjunctive particle	jj	
adnominal case particle	jm	
prefinal ending	ep	EP
final ending	ef	EF
conjunctive ending	ec	EC
nominalizing ending	etn	ETN
adnominalizing ending	etm	ETM

5 Experiments

5.1 Experimental environment

We used ETRI POS tagged corpus of 288,269 Eojoels for testing and the 21st Century Sejong Project's POS tagged corpus (Sejong corpus, for short) for training. The Sejong corpus consists of three different corpora acquired from 1999 to 2001. The Sejong corpus of 1999 consists of 1.5 million Eojeols and other two corpora have 2 million Eojeols respectively. The evaluation measures for the noun extraction task are recall, precision, and F-measure. They measure the performance by document and are averaged over all the test documents. This is because noun extractors are usually used in the fields of applications such as information retrieval (IR) and document categorization. We also consider the frequency of nouns; that is, if the noun frequency is not considered, a noun occurring twice or more in a document is treated as other nouns occurring once. From IR point of view, this takes into account of the fact that even if a noun is extracted just once as an index term, the document including the term can also be retrieved.

The performance considerably depends on the following factors: the representation schemes for word boundary detection, the tagset, the amount of training data, and the difference between training data and test data.

First, we compare four different representation schemes (BI, BIS, IE, IES) in word boundary detection as explained in Section 4. We try to use the following three kinds of tagsets in order to select the most optimal tagset through the experiments:

Tagset 1 Simply use two tags (e.g. noun and non-noun). This is intended to examine the syllable characteristics; that is, which syllables tend to belong to nouns or not.

Tagset 2 Use the tagset used in the training data without modification. ETRI tagset used for training is relatively smaller than that of other tagsets. This tagset is changeable according to the POS tagged corpus used in training.

Tagset 3 Use a simplified tagset for the purpose of noun extraction. This tagset is simplified by combining postpositions, adverbs, and verbal suffixes into one tag, respectively. This tagset is always fixed even in a different training corpus.

Tagset 2 used in Section 5.2 and Tagset 3 are represented in Table 2.

5.2 Experimental results with similar data

We divided the test data into ten parts. The performances of the model are measured by averaging over

Table 3: Experimental results of the ten-fold cross validation

	without considering frequency			with considering frequency		
	Precision	Recall	F-measure	Precision	Recall	F-measure
BI-1	72.37	83.61	77.58	74.61	82.47	78.34
BI-2	85.99	92.30	89.03	88.96	90.42	89.69
BI-3	84.85	91.20	87.90	87.56	89.55	88.54
BIS-1	78.50	83.53	80.93	80.36	83.99	82.13
BIS-2	**88.15**	**92.34**	**90.19**	**90.65**	**91.58**	**91.11**
BIS-3	86.92	91.07	88.94	89.27	90.62	89.94
IE-1	73.21	81.38	77.07	75.11	81.04	77.96
IE-2	85.12	91.54	88.21	88.37	90.34	89.34
IE-3	83.28	89.70	86.37	86.54	88.80	87.65
IES-1	78.07	82.69	80.31	79.54	83.08	81.27
IES-2	87.30	92.18	89.67	90.05	91.48	90.76
IES-3	85.80	90.79	88.22	88.46	90.47	89.45

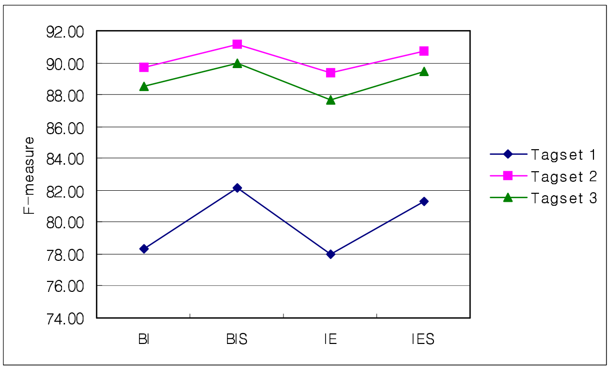

Figure 2: Changes of F-measure according to tagsets and representation schemes

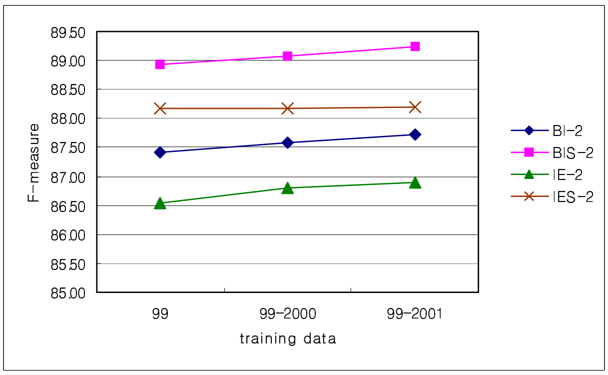

Figure 3: Changes of F-measure according to the size of training data

the ten test sets in the 10-fold cross-validation experiment. Table 3 shows experimental results according to each representation scheme and tagset. In the first column, each number denotes the tagset used. When it comes to the issue of frequency, the cases of considering frequency are better for precision but worse for recall, and better for F-measure. The representation schemes using single syllable information (e.g. "BIS", "IES") are better than other representation schemes (e.g. "BI", "IE"). Contrary to our expectation, the results of Tagset 2 consistently outperform other tagsets. The results of Tagset 1 are not as good as other tagsets because of the lack of the syntactic context. Nevertheless, the results reflect the usefulness of the syllable based processing. The changes of the F-measure according to the tagsets and the representation schemes reflecting frequency are shown in Figure 2.

5.3 Experimental results with different data

To show the influence of the difference between the training data and the test data, we have performed the experiments on the Sejong corpus as a training data and the entire ETRI corpus as a test data. Table 4 shows the experimental results on all of the three training data. Although more training data are used in this experiment, the results of Table 3 shows better outcomes. Like other POS tagging models, this indicates that our model is dependent on the text domain.

Table 4: Experimental results of Sejong corpus (from 1999 to 2001)

	without considering frequency			with considering frequency		
	Precision	Recall	F-measure	Precision	Recall	F-measure
BI-1	71.91	83.92	77.45	73.57	82.95	77.98
BI-2	85.38	89.96	87.61	87.19	88.26	87.72
BI-3	83.36	89.17	86.17	85.12	87.39	86.24
BIS-1	76.77	82.60	79.58	78.40	83.16	80.71
BIS-2	**87.66**	**90.41**	**89.01**	**88.75**	**89.75**	**89.25**
BIS-3	86.02	88.89	87.43	87.10	88.41	87.75
IE-1	70.82	79.97	75.12	72.67	79.64	75.99
IE-2	84.18	89.23	86.63	85.99	87.83	86.90
IE-3	82.01	87.67	84.74	83.79	86.57	85.16
IES-1	76.19	81.84	78.91	77.31	82.32	79.74
IES-2	86.41	89.33	87.85	87.66	88.75	88.20
IES-3	84.45	88.28	86.33	85.89	87.96	86.91

Table 5: Performances of other systems

	without considering frequency			with considering frequency		
	Precision	Recall	F-measure	Precision	Recall	F-measure
NE2001	84.08	91.34	87.56	87.02	89.86	88.42
KOMA	60.10	**93.12**	73.06	58.07	**93.67**	71.70
HanTag	**90.54**	88.68	**89.60**	**91.77**	88.58	**90.15**

Figure 3 shows the changes of the F-measure according to the size of the training data. In this figure, "99-2000" means 1999 corpus and 2000 corpus are used, and "99-2001" means all corpora are used as the training data. The more training data are used, the better performance we obtained. However, the improvement is insignificant in considering the amount of increase of the training data.

Results reported by Lee et al. (2001) are presented in Table 5. The experiments were performed on the same condition as that of our experiments. NE2001, which is a system designed only to extract nouns, improves efficiency of the general morphological analyzer by using positive and negative information about occurrences of nouns. KOMA (Lee et al., 1999b) is a general-purpose morphological analyzer. HanTag (Kim et al., 1998) is a POS tagger, which takes the result of KOMA as input. According to Table 5, HanTag, which is a POS tagger, is an optimal tool in performing noun extraction in terms of the precision and the F-measure. Although the best performance of our proposed model (BIS-2) is worse than HanTag, it is better than NE2001 and KOMA.

5.4 Limitation

As mentioned earlier, we assume that morphological variations do not occur at any inflected words. However, some exceptions might occur in a colloquial text. For example, the lexical level forms of two Eojeols "때(ddai)+는(neun)" and "고개(go-gai)+를(leul)" are changed into the surface level forms by contractions such as "땐(ddain)" and "고갤(go-gail)", respectively. Our models alone cannot deal with these cases. Such exceptions, however, are very rare. [6] In these experiments, we do not perform any post-processing step to deal with such exceptions.

6 Conclusion

We have presented a word recognition model for extracting nouns. While the previous noun extraction

[6] Actually, about 0.145% of nouns in the test data belong to these cases.

methods require morphological analysis or POS tagging, our noun extraction method only uses the syllable information without using any additional morphological analyzer. This means that our method does not require any dictionary or linguistic knowledge. Therefore, without manual labor to construct and maintain those resources, our method can extract nouns by using only the statistics, which can be automatically extracted from a POS tagged corpus.

The previous noun extraction methods take a morpheme as a processing unit, but we take a new notion of word as a processing unit by considering the fact that nouns belong to uninflected morphemes in Korean. By virtue of the new definition of a word, we need not consider mismatches between the surface level form and the lexical level one in recognizing words.

We have performed various experiments with a wide range of variables influencing the performance such as the representation schemes for the word boundary detection, the tag set, the amount of training data, and the difference between the training data and the test data. Without morphological analysis or POS tagging, the proposed method achieves comparable performance compared with the previous ones. In the future, we plan to extend the context to improve the performance.

Although the word recognition model is designed to extract nouns in this paper, the model itself is meaningful and it can be applied to other fields such as language modeling and automatic word spacing. Furthermore, our study make some contributions in the area of POS tagging research.

References

D.-U. An. 1999. A noun extractor using connectivity information. In *Proceedings of the Morphological Analyzer and Tagger Evaluation Contest (MATEC 99)*, pages 173–178.

S.-S. Kang and C.-W. Woo. 2001. Automatic segmentation of words using syllable bigram statistics. In *Proceedings of the 6th Natural Language Processing Pacific Rim Symposium*, pages 729–732.

S.-S. Kang. 1995. Morphological analysis of Korean irregular verbs using syllable characteristics. *Journal of the Korea Information Science Society*, 22(10):1480–1487.

N.-C. Kim and Y.-H. Seo. 1999. A Korean morphological analyzer CBKMA and a index word extractor CBKMA/IX. In *Proceedings of the MATEC 99*, pages 50–59.

J.-D. Kim, H.-S. Lim, S.-Z. Lee, and H.-C. Rim. 1998. Twoply hidden Markov model: A Korean pos tagging model based on morpheme-unit with word-unit context. *Computer Processing of Oriental Languages*, 11(3):277–290.

O.-W. Kwon, M.-Y. Chung, D.-W. Ryu, M.-K. Lee, and J.-H. Lee. 1999. Korean morphological analyzer and part-of-speech tagger based on CYK algorithm using syllable information. In *Proceedings of the MATEC 99*.

J.-Y. Lee, B.-H. Shin, K.-J. Lee, J.-E. Kim, and S.-G. Ahn. 1999a. Noun extractor based on a multipurpose Korean morphological engine implemented with COM. In *Proceedings of the MATEC 99*, pages 167–172.

S.-Z. Lee, B.-R. Park, J.-D. Kim, W.-H. Ryu, D.-G. Lee, and H.-C. Rim. 1999b. A predictive morphological analyzer, a part-of-speech tagger based on joint independence model, and a fast noun extractor. In *Proceedings of the MATEC 99*, pages 145–150.

D.-G. Lee, S.-Z. Lee, and H.-C. Rim. 2001. An efficient method for Korean noun extraction using noun occurrence characteristics. In *Proceedings of the 6th Natural Language Processing Pacific Rim Symposium*, pages 237–244.

D.-G. Lee, S.-Z. Lee, H.-C. Rim, and H.-S. Lim. 2002. Automatic word spacing using hidden Markov model for refining Korean text corpora. In *Proceedings of the 3rd Workshop on Asian Language Resources and International Standardization*, pages 51–57.

H.-S. Lim, S.-Z. Lee, and H.-C. Rim. 1995. An efficient Korean mophological analysis using exclusive information. In *Proceedings of the 1995 International Conference on Computer Processing of Oriental Languages*, pages 225–258.

Lance A. Ramshaw and Mitchell P. Marcus. 1995. Text chunking using transformation-based learning. In *Proceedings of the Third Workshop on Very Large Corpora*, pages 82–94.

J.-H. Shim, J.-S. Kim, J.-W. Cha, and G.-B. Lee. 1999. Robust part-of-speech tagger using statistical and rule-based approach. In *Proceedings of the MATEC 99*, pages 60–75.

K.-S. Shim. 1996. Automated word-segmentation for Korean using mutual information of syllables. *Journal of the Korea Information Science Society*, 23(9):991–1000.