A Hidden Markov Model Imbedding Multiword Units for Part-of-Speech Tagging

Jae-Hoon Kim and Jungyun Seo

Abstract

Morphological Analysis of Korean has known to be a very complicated problem. Especially, the degree of part-of-speech(POS) ambiguity is much higher than English. Many researchers have tried to use a hidden Markov model(HMM) to solve the POS tagging problem and showed arround 95% correctness ratio. However, the lack of lexical information involves a hidden Markov model for POS tagging in lots of difficulties in improving the performance. To alleviate the burden, this paper proposes a method for combining multiword units, which are types of lexical information, into a hidden Markov model for POS tagging. This paper also proposes a method for extracting multiword units from POS tagged corpus. In this paper, a multiword unit is defined as a unit which consists of more than one word. We found that these multiword units are the major source of POS tagging errors. Our experiment shows that the error reduction rate of the proposed method is about 13%.

I. Introduction

Part-of-speech (hereafter POS) tagging is to assign a POS to each word in a sentence. The POS tagging system is widely used in speech recognition and synthesis, information retrieval as well as natural language processing. The accuracy of most of them is at least 95%, with practically no restrictions on the input text 5]. This means that there is one tagging error in every 20 words tagged by the system. Tagging errors may cause serious problems in most applications. In the case of syntactic parsing, a tagging error a causes parsing error or failure. This is the motivation of this research: improving the performance of the POS tagging using all possible information. A hidden Markov model (hereafter, HMM) is well-known technique for POS tagging. In the model, one of its problems is that it is not easy to reflect lexical contextual information (hereafter, LCI) although the LCI plays an important role in POS tagging[8, 16]. For example, the POS of the word, '*sound*', is a noun in the sentence '*sound energy*', an adjective in the phrase '*sound fruit*', and a verb in the phrase "*They sound alarms.*" As we can see in these examples, the POSs of some words are affected by the surrounding words rather than surrounding POSs. We propose a method for reflecting the LCI on an HMM to improve the performance of the tagging system. To model the LCI on the HMM, we should solve two problems: One is how to combine the LCI such as multiword units into an HMM, and the other is how to determine the effective LCIs. We have slightly modified the HMM for the former problem and introduced a method to extract effective collocations for the latter problem. The proposed method has reduced the error rate by about 13% as compared with the original HMM. We expect that the proposed method would show more promising result if we manually elaborate the LCI. However, since manually extracting the LCI is labor intensive, it should be automatically extracted as we suggested in this paper.

This paper is organized as follows; In Section II, we discuss HMM and Korean POS tagging as background works. In Section III and IV, combination of an HMM and multiword units as an LCI, and the method for extracting the multiword units are described, respectively. After presenting some experimental results and comparing with other works in Section V and VI, we summarize our findings and draw conclusions in Section VII.

Manuscript received September 15, 1997; accepted November 20, 1997.

J. H. Kim is with Division of Automation and Information Engineering, Korea Maritime University, Pusan, Korea.

J. Y. Seo is with Department of Computer Science, Sogang University, Seoul, Korea.

II. Background

1. HMM for POS Tagging

We introduce an HMM, a well known probabilistic model for POS tagging. In the model, a POS tagging procedure ϕ is to select the most proper POS sequence T_i satisfying with Equation (1) in a given sentence W[1, 3, 9].

$$\phi(W) \equiv \operatorname{argmax}_{T_i} \Pr(T_i \mid W) = \operatorname{argmax}_{T_i} \Pr(T_i, W) \quad (1)$$

Equation (2) is derived from Equation (1) by using the Markov assumption and the chain rule, where the input sentence W is $w_1, w_2, \ldots, w_n$ and the most proper POS sequence for W is $t_1, t_2, \ldots, t_n$.

$$\phi(W) = \operatorname{argmax}_T \prod_{i=1}^{n} \Pr(t_i \mid t_{i-2}, t_{i-1}) \Pr(w_i \mid t_i) \quad (2)$$

This equation is called the second order HMM for POS tagging. On the right side of Equation (2), the first part is called a contextual probability, and the second a lexical generation probability[17].

2. *Korean POS Tagging*

Korean is different from English in word-formation as well as word order. According to the differences, the definition of POS tagging can vary slightly. English POS tagging assigns the most proper POS to each word in a given sentence as mentioned in Section 1[1, 17]. On the other hand, Korean POS tagging assigns not only the most proper sequence POSs but also the most proper sequence of morphemes to each Eojeol[1] in a given sentence [9][2]. We widely use a well-known HMM for Korean POS tagging like English POS tagging. According to whether the information is included between Eojeols or not, Korean POS tagging can be classified into two models, an Eojeol-based POS tagging model [10, 12] and a morpheme-based POS tagging model[8, 9, 11]. In the former, a tag of an Eojeol is represented as the POS sequence of morphemes[12] or a POS pair which is the beginning and the end of the POS sequence of morphemes[10] for a given Eojeol. An advantage of this model is to consider the contextual information for Eojeols as well as morphemes. On the other hand, a disadvantage is not to fix the number of Eojeol tags, therefore data sparseness and some Eojeol ambiguities on the same POS sequence arise. In the latter, the number of morpheme tags is fixed and small, but the contextual information for Eojeol can not be reflected. Recently, to improve the performance, a hybrid model begins to appear on the stage of Korean POS tagging. As a representative example of the hybrid model, there is a model that is combined with HMM and rules like Brill's transformation [14, 15]

III. Combination of Multiwords and HMM

In this paper, a multiword unit is defined as a unit which consists of more than one adjacent word without considering to a grammatical unit in a sentence. Of course, most grammatical units consisting of more than one word belong to multiword units.

Table 1 shows some examples of multiword units[3]. In Table 1,

Table 1. Some examples of multiword units.

No.	Multiword unit (Meaning)	Remarks
1	'*hanpich cwunghakkyo*' (Hanbit Middle School)	a proper noun
2	'*cenhwa penho*' (telephone number)	a compoud noun
3	'*-ko iss-*' (be -ing)	an auxiliary conjunctive ending and an auxiliary verb
4	'*-ey kwanha-*' (with regard to)	a particle and a verb
5	'kkamccak nolla-' (be startled all of sudden)	a adverb and a verb.
6	'*kolthang mek-*' (be cheated)	a noun and a verb

1 and 2 are grammatical units. 3 and 4 are a functional word and a content word, which are closely related together. 5 and 6 are some collocative words like '*take place*' in English. Except for those in Table 1, there are several sorts of multiword units as in Table 5. In combining multiword units into an HMM, we should solve two problems: One is how to include multiword units in an HMM without changing the original model, and the other is how to extract multiword units from texts or corpus. We will describe the problems in the following subsections.

1. *A Multiword Unit Based POS Tagging Model*

A multiword-based POS tagging model, which is based on the h-order HMM, is defined by

$$\phi(W) \simeq \operatorname{argmax}_{t_{1,n}} \prod_{i=1}^{n} \begin{cases} \Pr(t_i \mid t_{i-h,i-1}) \Pr(w_{i-k,i} \mid t_i) & \text{if } \phi = \mathrm{m}_{i-1} \cap \mathrm{m}_i \\ \dfrac{\Pr(t_i \mid t_{i-h,i-1})}{\Pr(t_{i-1} \mid t_{i-h,i-2})} \dfrac{\Pr(w_{i-k,i} \mid t_{i-k,i})}{\Pr(w_{i-p,i} \mid t_{i-p,i})} & \text{otherwise,} \end{cases} \quad (3)$$

where W is an input sentence, T is a correct POS sequence for W, k is the length of a multiword unit minus one, and p is the length of intersected words between a previous multiword unit and a current multiword unit. $w_{1,n}(=w_1 w_2 \ldots w_n)$ denotes a sentence with n words. In a similar way, $t_{1,n}(=t_1 t_2 \ldots t_n)$ denotes a POS sequence for $w_{1,n}$. Next we want to probe relations between parameters h, k, and p of Equation (3). $k(0 \le k < h)$ has a different value from state to state. For the original HMM, the values of k on all states are zero. Now we consider the value of k to be one. The multiword unit $w_{i-1,i}$ is denoted by m_i and consists of two words, w_{i-1} and w_i. Therefore, an observation symbol on each state is a word or two words according to the

1) An Eojeol is a sequence of morphemes between two spaces and is very similar to a word in English.

2) Note that readers can find the other differences in another paper of author [9], but not mentioned in this paper.

3) In this paper, the Yale Romanization is used to represent Korean words and sentences.

Table 2. The structures of some Eojeols.

Eojeols	*hakkyoey*	*kako*	*issta.*
Morphological structures	*hakkyo*/nc+*ey*/jca	*ka*/pv+*ko*/ecq	*iss*/pa+*ta*/ef+./s.
		kal/pv+*ko*/ecq	*iss*/px+*ta*/ef+./s.
		ka/px+*ko*/ecq	
		ka/pv+*ko*/ecx	
		kal/pv+*ko*/ecx	
		ka/px+*ko*/ecx	

value of k in case of $h = 2$. p is the length of intersected words as mentioned above and the value of p is $0 \le p \le k$. In Equation (3), the above equation without intersected words and the numerator of the below equation are the same with the original HMM. The denominator of the below equation, however, prevents the probabilities of the multiword units from reflecting the intersected words on the final sequence twice. In Figure 1, as an example, we explain details about this model with the help of Table 2, which shows the morphological structures of Eojeols in a Korean sentence "*hakkyoey kako issta*(I go to school)."[4] The figure shows a (weighted) network (lattice) of the example sentence as an observation sequence based on the second order HMM. n the figure, a state and a transition of the HMM are represented by a node and an edge, respectively, and an observation symbol is labeled on the right and top of each node. Thus, the values on a node and an edge, but disappeared from the figure, mean a (multiword unit-based) lexical probability and a (multiword unit-based) contextual probability, respectively. The most proper sequence is represented by a bold line and \$ is a special symbol to represent the beginning and the end of a sentence in the figure. To help readers to understand this model fully, we explain this model in detail through a concrete example. Suppose that h be 2 for convenience. So, k can be 0 or 1. For a given Korean sentence "*hakkyoey kako issta.*" which is the same sentence given in the example in Figure 1, the valid morphological analysis is "*hakky* / nc + *ey*/jca *ka* / pv + *ko* / ecx *iss* / px + *ta* / ef + . / s." which is underlined in Table 2. Suppose that '*ko iss-*' be a multiword unit with $k = 1$. All words except this word are simple words (morphemes) with $k = 0$. Then, the probability Pr (*hakkyo+ey ka+ko iss+ta+*., nc+jca pv+ecx px+ef+s.) of Equation (3) is calculated by the followings;

$$\begin{aligned}&\Pr(hakkyo \mid nc)\ \Pr(nc \mid \$,\ \$) \times \Pr(ey \mid jca)\ \Pr(jca \mid \$,\ nc) \times \\ &\Pr(ka \mid pv)\ \Pr(pv \mid nc,\ jca) \times \Pr(ko \mid ecx)\ \Pr(ecx \mid jca,\ pv) \times \\ &\frac{\Pr(ko, iss \mid ecx, px)\ \Pr(ecx, px \mid pv)}{\Pr(ko \mid ecx)\ \Pr(ecx \mid pv)} \times \\ &\Pr(ta \mid ef)\ \Pr(ef \mid ecx,\ px) \times \Pr(. \mid s.)\ \Pr(s. \mid pv,\ ef) \times \\ &\Pr(\$ \mid \$)\ \Pr(\$ \mid ef,\ s.)\end{aligned} \quad (4)$$

4) A morphological structure is a result of morphological analysis for an Eojeol and is regarded as a linear structure of which elements are distinguished by a delimiter '+'. '*hakkyo*/nc' means that the POS tag of the morpheme '*hakkyo*' are 'nc'. We put the list of Korean POS tags in an appendix.

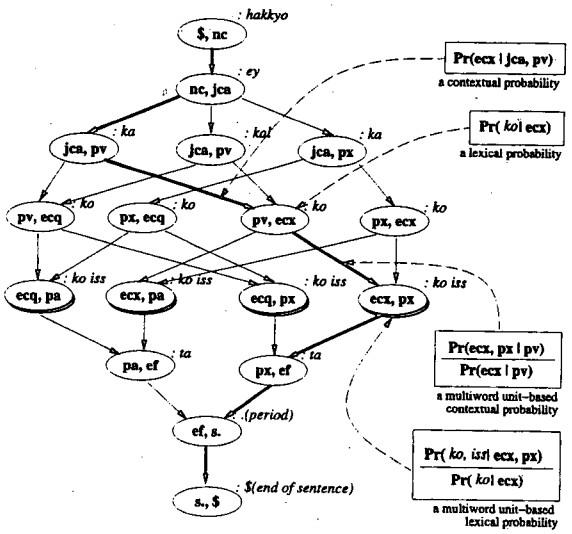

Fig. 1. A weighted network (lattice) of observations and states (nodes) based on the second order HMM.

2. *Parameter Estimation of the Multiword Units Based POS Tagging Model*

If k is 0, parameter estimation is the same as the original HMM. Now we turn to parameter estimation in case of $k = 1$. Consider a special case as an example in case of $k = 1$ and $h = 2$ for our experiment described below. Lexical probability and contextual probability for a multiword unit $w_{i-1,i}$ are estimated by Equation (5) and (6);

$$\Pr(w_{i-1,i} \mid t_{i-1,i}) \simeq \frac{C(w_{i-1,i}, t_{i-1,i})}{C(t_{i-1,i})} \quad (5)$$

$$\Pr(t_{i-1,i} \mid t_{i-2}) \simeq \begin{cases} \dfrac{C(t_{i-2,i})}{C(t_{i-2})} & \text{if } k = 1 \\ \dfrac{C(t_{i-2,i})}{C(t_{i-2,i-1})} & \text{otherwise}(k = 0), \end{cases} \quad (6)$$

where $C(x)$ is a frequency of x. 3-gram for POSs is sufficient to estimate parameters for contextual probability in the second order HMM and $k = 1$ as you can see in Equation (6). Therefore, the degree of the data sparseness is the same with the original HMM. In proportion to using some kinds of multiword units, however, the lexical probability is very different from the contextual probability in data sparseness, which might cause the performance to make worse. To alleviate this problem, we use multiword units including an error-prone word with high frequency. We will describe the extraction method of multiword units in the next section in detail.

3. *Extraction of Multiword Units*

The extraction method of multiword units is similar to that of collocation in respect of using the frequency of n-gram[13]. This is the difference in that the frequency is counted in not all words, but only in error-prone words. That is, if w_i is an error-prone word,

the frequency of multiword units including w_i is defined by Equation (7);

$$C_e(w_{i-k,i}) \text{ or } C_e(w_{i,i+k}) > \rho_1, \quad (7)$$

where w_i is an error-prone word, $C_e(w_{i-k,i})$ and $C_e(w_{i,i+k})$ are the frequency of the left context and the right context of the error-prone word w_i, respectively, and $\rho_1(\rho_1 > 1)$ is a constant as a threshold. Generally, mutual information is used for extracting collocation[4], but is improper in extracting multiword units. This is the reason that the left and the right context are considered differently. Let us consider the left context of an error-prone Korean word '*-ko*'. The left context can be all verbs such as '*mek*(eat) + *ko*' and '*ip*(wear) + *ko*' *etc.* In this paper, these verbs are improper as multiword units based on '*-ko*'. On the other hand, consider the right context of the word '*-ko*'. In many cases, the right context is a special auxiliary verb '*iss-*', but this might not always be the case. Therefore, in the case of the error-prone word '*-ko*', the left context is not proper as a multiword unit, but the right context is proper. In this paper, as we pay attention to this point, the conditional probability and the relative frequency count[21] are used for extracting multiword units as in Equation (8).

$$\Pr(w_{i-k,i-1}|w_i)=\frac{C(w_{i-k,i})}{\mathrm{E}\{C(w_{i-k,i})\}} \text{ or } \Pr(w_{i+1,i-k}|w_i)\frac{C(w_{i,i-k})}{\mathrm{E}\{C(w_{i,i-k})\}} > \rho_2, \quad (8)$$

where $\mathrm{E}\{C(w_{i-k,i})\}$ and $\mathrm{E}\{C(w_{i,i+k})\}$ are the average frequency of $w_{i-k,i}$ and $w_{i,i+k}$, respectively, $\rho_2(\rho_2 > 0)$ is a constant as a threshold, w_i is an error-prone word. In this paper, ρ_1 and ρ_2 are controlled to keep minimal errors on the training corpus described in next section in detail.

IV. Experiments and Evaluation

The main objective of this paper is to show that the multiword unit is a kind of useful information to improve the performance of a tagging system, especially based on an HMM.

1. Experimental Environment

We use the "KAIST corpus" data described in Kim (1996). It contains 15,950 sentences and its other statistics are shown in Table 3. These sentences have been tagged manually at the department of computer science in KAIST. The training corpus and the test corpus are independent. We use 51 different POS tags as in Appendix. We have built a dictionary that indicates the list of possible tags for each morpheme, by taking all the words that occur in the total corpus. In similar way, we have established a multiword unit dictionary by using the extraction method described in Section 3.3. Thus, these are a closed dictionary since a word will not have all its possible tags although the tags actually are within the corpus. In Korean, a morphological analyzer plays an important role in POS tagging. We used the morphological analyzer based on lexicalized morphotactics[8] for our experiment.

Table 3. Statistics of training and test corpus.

statistics	training corpus	test corpus
no. of sentences	12,082	3,868
no. of Eojeols	131,581	41,122
no. of morphemes	284,241	88,683
avg. no. of Eojeols per sentence	10.89	10.63
avg. no. of morphemes per Eojeol	2.16	2.16

Table 4. Performance according to model parameters.

no. multiword units	ρ_1	ρ_2	no. of errors		for morphemes
			Eojeol	morphemes	error reduction rate(%)
50	-	-	1655	1987	0.00
29	5	3.000	1606	1889	4.93
43	5	1.000	1601	1885	5.13
60	3	1.000	1589	1876	5.59
78	2	1.000	1591	1878	5.49
120	3	0.100	1533	1786	10.12
130	3	0.050	1507	1749	11.98
143	**3**	**0.010**	**1493**	**1733**	**12.78**
146	3	0.005	1493	1733	12.78
151	3	0.001	1495	1737	12.58

2. Performance Evaluation

In this experiment, we extracted 3-gram of POS from the training corpus. Then, we computed the relative frequency count as the supervised parameter estimation method and used the Good-turning method[6] for smoothing. This model was then used to tag the test sentence in the test corpus. The results are indicated in Table 4. The table shows that the performance varies as the control of two model parameters, ρ_1 and ρ_2. Note that the first row on the table is the performance concerned in the second order original HMM. In our experiment, the number of selected multiword units is determined according to the value of ρ_1 and ρ_2 in the training corpus. We get the best result in the case of $\rho_1=3$ and $\rho_2=0.01$. As a result, the error reduction rate is about 13%. Total tagging accuracy is about 98%: about 0.2% improved.

3. Selected Multiword Units

In our experiment, Table 5 shows a part of the selected multiword units of which some are not intuitive. In the table, the functional words are marked with an asterisk '*'. A selected multiword unit has at least one functional word. This means that most error-prone words are functional words in Korean.

A great number of endings are especially error-prone functional words. The determination of correct POS for the endings requires

Table 5. A part of the extracted multiword units in our experiment.

Left context of error-prone word		Right context of error-prone word	
(w_{i-1}	w_i)	(w_i	w_{i+1})
*i**	*ta**	*ha*	*n**
ha	*e**	*ke**	*i**
*ass**	*ta**	*ha*	*e**
*i**	*nka**	*key**	*twi*
ha	*ko**	*ha*	*ess**
ha	*key**	*ha*	*ko**
*key**	*twi*	*ko**	
*nun**	*ke**	*ha*	*nun**
*e**	*poi*	*il*	*i**
*e**	*naka*	*sulep**	*n**
*i**	*ya**	*ha*	*l**
ha	*mye**	*twi*	*ess**
*i**	*lan**	*ha*	*nta**
*n**	*il*	*twi*	*e**
*ey**	*ilu*	*m**	*ul**
*ul**	*tut*	*yeph*	*ey**
*lul**	*tut*	*key**	*ha*

syntactic analysis, but it is somewhat, but not completely, resolved by observing some words around the error-prone endings. A representative example is a phrase constituted by an auxiliary conjunctive ending and an auxiliary verb.

V. Discussions

For POS tagging, a VMM (variable Memory Markov) model proposed by Schütze and Singer (1994) is similar in using variable-length context to our method. Both methods also adjust the length of context using errors. In order to determine the context, Schütze and Singer use the statistical error based on relative entropy, while we use the error environment including at least one error-prone word based on the conditional probability and relative frequency count. Another difference is a type of variable contexts, that is, they use only POSs while we use LCIs as well as POSs. Brill's method[2] can also accept variable contexts. It, furthermore, have the nature of long-distance correlations as well, but our proposed methods neglect it due to the Markov nature. This is a drawback of our proposed methods. There is another tagging model with variable context, which is called PCM (probabilistic classification model) proposed by Lin, Chiang, and Su (1994). PCM is also similar to our proposed method in applying to error-prone words. PCM re-tags POSs to error prone words selected by CART while our method do not. Now we turn to a method for extraxting multiword units, which is very similar to that for extracting collocations[4, 20, 21]. Especially our approach is similar in using relative frequency count to the approach proposed by Su, Wu, and Chang(1994). We, however, use the conditional probability as mentioned in Section III.3. We observe that the conditional probability is good for extracting the selectional restrictions through another experiment[13].

VI. Concluding Remarks

In this paper, we have presented a POS tagging model with combining multiword units into an HMM and a method for extracting multiword units from POS tagged corpus. In this paper, the multiword units are defined as more than one word which frequently cause POS tagging errors. Our experiment shows an error reduction rate of about 13% as compared with the original HMM and a total accuracy of about 98%. The results of experiments reveal that multiword units are well-suited to a type of the lexical contextual information on an HMM. We expect that the proposed method shows the more promising results if multiword units (not selected automatically, but error-prone words explicitly) could be added manually, but laboriously, rather than automatically.

Acknowledgement

This work was supported in part by the Ministry of Information and Communication under the title of "A Research on Multimodal Dialogue Interface".

References

[1] Allen, J. *Natural Language Understanding*, 2nd edition, The Benjamin/ Cummings Publishing Company, Inc.

[2] Brill, E. (1995), "Transformation-based error driven learning and natural language processing: a case study in part-of-speech tagging", *Computational Linguistics*, Vol. 21, No. 4, pp. 543-564.

[3] Charniak, E., Hendrickson, C., Jacobson, N., and Perkowitz, M. (1993), "Equations for part-of-speech tagging", *Proceedings of National Conference on Artificial Intelligence (AAAI-93)*, pp. 748-789.

[4] Church, K. W. and Hanks, P. (1990), "Word association norms, mutual information, and lexicography", *Computational Linguistics*, Vol. 16, No. 1, pp. 22-29.

[5] Church, K. W. and Mercer, R. L. (1993), "Introduction to the special issue on computational linguistics using large corpora", *Computational Linguistics*, Vol. 19, No. 1, pp. 1-24.

[6] Good, I. J. (1953), "The population frequencies of species and the estimation of population parameters", *Biometrika*, Vol. 40, No. 3, pp. 237-264.

[7] Kim, J.-D., Lim, H.-S., and Rim, H.-C. (1996), "A Korean part-of-speech tagging model based on morpheme unit with Eojeol-unit context", *Proceedings of the Korea Cognitive Science Society Spring Conference*, pp. 97-106(in Korean).

[8] Kim, J.-H. (1996), *Lexical Disambiguation Using Error-Driven Learning*, Ph.D. Thesis, Dept. of Computer Science, KAIST (in Korean).

[9] Kim, J.-H., Lim, C.-S., and Seo, J. (1995), "An efficient Korean part-of-speech tagging using hidden Markov model", *Journal of the Korea Information Science Society*, Vol. 22, No. 1, pp. 136-146(in Korean).

[10] Lee, H.-K. (1997), "An effective estimation method for lexical probabilities in Korean lexical disambiguation", *Journal of the Korea Information Processing Society*, Vol. 3, No. 6, pp. 1588-1597(in Korean).

[11] Lee, S.-H. (1995), *A Korean Part-of-Speech Tagging System with Handling Unknown Words*, M.S. Thesis, Dept. of Computer Science, KAIST(in Korean).

[12] Lee, W.-J. (1993), *Design and Implementation of an Automatic Tagging System for Korean Texts*, M.S. Thesis, Dept. of Computer Science, KAIST(in Korean).

[13] Lee, K. J., Kim, J.-H., and Kim, G. C. (1995), "Extracting collocations from tagged corpus in Korean", *Proceedings of the 22nd KISS Spring Conference*, Inha Univ., Incheon, pp. 623-626(in Korean).

[14] Lee, J.-H. and Shin, S.-H (1995), "TAKTAG: Two phase learning method for hybrid statistical/rule-based part-of-speech disambiguation", *Proceedings of the 6th International Conference on Computer Processing of Oriental Languages (ICCPOL-95)*, Hawaii.

[15] Lim, H.-S., Kim, J.-D, and Rim, H.-C. (1997), "A Korean part-of-speech tagger using transformation-based error-driven learning", *Proceedings of the 7th International Conference on Computer Processing of Oriental Languages(ICCPOL-97)*, Hong Kong Baptist Univ. pp. 456-459.

[16] Lin, Y.-C., Chiang, T.-H., and Su, K.-Y. (1994), "Automatic model refinement - with an application to tagging", *Proceedings of International Conference on Computational Linguistics(COLING-94)*, Kyoto, Japan, pp. 148-153.

[17] Merialdo, B. (1994), "Tagging English text with a probabilistic model," Computational Linguistics", Vol. 20, No. 2, pp. 156-171.

[18] Shin, J.-H., Han, Y. S., Park, Y. C., and Choi, K.-S., (1994) "A HMM part-of-speech tagger for Korean with word phrasal relations", *Proceedings of International Conference on Recent Advances in Natural Language Processing (RANLP-94)*, Sophia, Bulgaria.

[19] Schütze H. and Y. Singer (1994), "Part-of-speech tagging using a variable memory Markov model", *Proceedings of the 26th Annual Meeting of the Assoc. for Computational Linguistics (ACL-32)*, pp. 181-187.

[20] Smadja, F. (1993), "Retrieving Collocations from Text: Xtract", *Computational Linguistics*, Vol. 19, No. 1, pp. 143-177.

[21] Su, K.-Y., Wu, M.-W., and Chang, J.-S. (1994), "A corpus-based approach to automatic compound extraction", *Proceedings of the 32th Annual Meeting of the Assoc. for Computational Linguistics(ACL-94)*}, pp. 242-247.

Appendix: Korean POS tags

1	s,	comma	2	s.	sentence closer
3	s‘	left quotation and parenthesis mark	4	s’	right quatation and parenthesis mark
5	s-	connection mark	6	su	unit
7	sw	currency	8	sy	other symbols
9	f	foreign word	10	nca	active common noun
11	ncs	stative common noun	12	nc	common noun
13	nq	proper noun	14	nbu	unit bound noun
15	nb	bound noun	16	npp	personal pronoun
17	npd	demostrative pronoun	18	nnn	number
19	nn	numeral	20	pv	verb
21	pad	demonstrative adjective	22	pa	adjective
23	px	auxiliary verb	24	md	demonstrative adnoun
25	mn	numeral adnoun	26	m	adnoun
27	ad	demonstrative adverb	28	ajw	word-conjunctive adverb
29	ajs	sentence-conjunctive adverb	30	a	adverb
31	i	interjection	32	jc	case particle
33	jcm	adnominal case particle	34	jcv	vocative case particle
35	jca	adverbial case partcle	36	jcp	predicative case particle
37	jx	auxiliary particle	38	jj	conjunctive particle
39	ecq	equal conjunctive ending	40	ecs	subordinative conjunctive ending
41	ecx	auxiliary conjunctive ending	42	exm	adnominal ending
43	exn	nominal ending	44	exa	adverbial ending
45	efp	prefinal ending	46	ef	final ending
47	xn	noun suffix	48	xpv	verb-derived suffix
49	xpa	adjective-derived suffix	50	xa	adverb-derived suffix
51	sp	a space, specical tag			

Jae-Hoon Kim is a faculty member of the Department of Computer Engineering, Korea Maritime University in Pusan, Korea. He had worked at Electronics and Telecommunications Research Institute(ERTI), Taejon, Korea as a seninor member of research staff. His research interests include corpus representation for information encoding, dialogue machine translation, natural (written/spoken) language processing, especially, part-of-speech tagging and statistical parsing. He obtained a B.A. in computer science from Keimyung University, Taegu, Korea, and an M.S and Ph.D. in computer science from KAIST, Taejon, Korea.

Jungyun Seo is an associate professor at the Department of Computer Science, Sogang University in Seoul, Korea. Previously, he was an assistant professor at the Computer Science Department at the Korea Advanced Institute of Science and Technology(KAIST) in Taejon, Korea. He had worked at the UniSQL, Inc., Austin, Texas as a member of technical staff. His research interests include Natural Language Processing, Computer-Human Interface, especially, Multi-Modal Dialogue interface. Seo obtained a BA in mathematics from Sogang University, Seoul, Korea, and an MS and PhD in computer science from the University of Texas, Austin.