Generalized unknown morpheme guessing for hybrid POS tagging of Korean*

Jeongwon Cha and Geunbae Lee and Jong-Hyeok Lee
Department of Computer Science & Engineering
Pohang University of Science & Technology
Pohang, Korea
{himen, gblee, jhlee}@postech.ac.kr

Abstract

Most of errors in Korean morphological analysis and POS (Part-of-Speech) tagging are caused by unknown morphemes. This paper presents a generalized unknown morpheme handling method with POSTAG(POStech TAGger) which is a statistical/rule based hybrid POS tagging system. The generalized unknown morpheme guessing is based on a combination of a morpheme pattern dictionary which encodes general lexical patterns of Korean morphemes with a posteriori syllable tri-gram estimation. The syllable tri-grams help to calculate lexical probabilities of the unknown morphemes and are utilized to search the best tagging result. In our scheme, we can guess the POS's of unknown morphemes regardless of their numbers and positions in an eojeol, which was not possible before in Korean tagging systems. In a series of experiments using three different domain corpora, we can achieve 97% tagging accuracy regardless of many unknown morphemes in test corpora.

1 Introduction

Part-of-speech (POS) tagging has many difficult problems to attack such as insufficient training data, inherent POS ambiguities, and most seriously unknown words. Unknown words are ubiquitous in any application and cause major tagging failures in many cases. Since Korean is an agglutinative language, we have unknown morpheme problems instead of unknown words in our POS tagging.

The usual way of unknown-morpheme handling before was to guess possible POS's for an unknown-morpheme by checking connectable functional morphemes in the same eojeol[1] (Kang, 1993). In this way, they could guess possible POS's for a single unknown-morpheme only when it is positioned in the begining of an eojeol. If an eojeol contains more than one unknown-morphemes or if unknown-morphemes appear other than the first position, all the previous methods cannot efficiently estimate them. So, we propose a morpheme-pattern dictionary which enables us to treat unknown-morphemes in the same way as registered known morphemes, and thereby to guess them regardless of their numbers and positions in an eojeol. The unknown-morpheme handling using the morpheme-pattern dictionary is integrated into a hybrid POS disambiguation.

The POS disambiguation has usually been performed by statistical approaches mainly using hidden markov model (HMM) (Cutting et al., 1992; Kupiec, 1992; Weischedel et al., 1993). However, since statistical approaches take into account neighboring tags only within a limited window (usually two or three), sometimes the decision cannot cover all linguistic contexts necessary for POS disambiguation. Also the approaches are inappropriate for idiomatic expressions for which lexical terms need to be directly referenced. The statistical approaches are not enough especially for agglutinative languages (such as Korean) which have usually complex morphological structures. In agglutinative languages, a word (called eojeol in Korean) usually consists of separable single stem-morpheme plus one or more functional morphemes, and the POS tag should be assigned to each morpheme to cope with the complex morphological phenomena. Recently, rule-based approaches are

* This project was supported by KOSEF (teukjeongkicho #970-1020-301-3, 1997).

[1] An eojeol is a Korean spacing unit(similar to English word) which usually consists of one or more stem morphemes and functional morphemes.

re-studied to overcome the limitations of statistical approaches by learning symbolic tagging rules automatically from a corpus (Brill, 1992; Brill, 1994). Some systems even perform the POS tagging as part of a syntactic analysis process (Voutilainen, 1995). However, rule-based approaches alone, in general, are not very robust, and not portable enough to be adjusted to new tag sets and new languages. Also the performance is usually no better than the statistical counterparts (Brill, 1992). To gain the portability and robustness and also to overcome the limited coverage of statistical approaches, we adopt a hybrid method that can combine both statistical and rule-based approaches for POS disambiguation.

2 Linguistic characteristics of Korean

Korean is classified as an agglutinative language in which an eojeol consists of several number of morphemes that have clear-cut morpheme boundaries. For examples, "나는 감기에 걸렸다(I caught a cold)" consists of 3 eojeols and 7 morpheme, such as[2] 나(I)/T + 는(auxiliary particle)/jS, 감기(cold)/MC + 에(other particle)/jO, 걸리(catch)/DR + 었(past tense)/eGS + 다(final ending)/eGE. Below are the characteristics of Korean that must be considered for morphological-level natural language processing and POS tagging.

- As an agglutinative language, Korean POS tagging is usually performed on a morpheme basis rather than an eojeol basis. So, morphological analysis is essential to POS tagging because morpheme segmentation is much more important and difficult than POS assignment. Moreover, morphological analysis should segment out unknown morphemes as well as known morphemes, so unknown morpheme handling should be integrated into the morphological analysis process. There are three possible analyses from the eojeol "나는" : '나(I)/T' + '는(subject-marker)/jS', '나(sprout)/DR' + '는(adnominal)/eCNMG', '날(fly)/DI' + '는(adnominal)/eCNMG', so morpheme segmentation is often ambiguous.
- Korean is a postpositional language with many kind of noun-endings (particles), verb-endings (other endings), and prefinal verb-endings (prefinal endings). It is these functional morphemes, rather than eojeol's order, which determine most of the grammatical relations such as noun's syntactic functions, verb's tense, aspect, modals, and even modifying relations between eojeols. For example, '는/jS' is an auxiliary particle, so eojeol "나는" has a subject role due to the particle '는/jS'.
- Complex spelling changes frequently occur between morphemes when two morphemes combine to form an eojeol. These spelling changes make it difficult to segment the original morphemes out before assigning the POS tag symbols.

Fig. 1 shows a tag set extracted from 100 full POS tag hierarchies in Korean. This tag set will be used in our experiments in section 6.

3 Unknown morpheme guessing during morphological analysis

Morphological analysis is a basic step to natural language processing which segments input texts into morphotactically connectable morphemes and assigns all possible POS tags to each morpheme by looking up a morpheme dictionary.
Our morphological analysis follows general three steps (Sproat, 1992): morpheme segmentation, original morpheme recovery from spelling changes, and morphotactics modeling. Input texts are scanned from left to right, character[3]by character, to be matched to morphemes in a morpheme dictionary. The morpheme dictionary (Fig. 2) has a separate entry for each variant form (called allomorph) of the original morpheme form so we can easily reconstruct the original morphemes from spelling changes.
For morphotactics modeling, we used the POS tags and the morphotactic adjacency symbols in the dictionary. The full hierarchy of POS tags and morphotactic adjacency symbols are encoded in the morpheme dictionary for each mor-

[2]Here, '+' is a morpheme boundary in an eojeol and '/' is for the POS tag symbols (see Fig. 1).

[3]The character sequence in "나는" is 'ㄴ', 'ㅏ', 'ㄴ', 'ㅡ', 'ㄴ'.

tag	description	tag	description	tag	description
MC	common noun	MPN	person name	MPC	country name
MPP	place name	MPO	other proper noun	MD	bound noun
T	pronoun	G	adnoun	S	numeral
B	adverb	K	interjection	DR	regular verb
DI	irregular verb	HR	regular adjective	HI	irregular adjective
I	i-predicative particle	E	existential predicate	jC	case particle
jS	auxiliary particle	jO	other particle	eGE	final ending
eGS	prefinal ending	eCNDI	aux conj ending	eCNDC	quote conj ending
eCNMM	nominal ending	eCNMG	adnominal ending	eCNB	adverbial ending
eCC	conjunctive ending	y	predicative particle	b	auxiliary verb
+	prefix	-	suffix	su	unit symbol
so	other symbol	s'	left parenthesis	s'	right parenthesis
s.	sentence closer	s-	sentence connection	s,	sentence comma
sf	foreign word	sh	Chinese character		

Figure 1: A tag set with 41 tags from 100 full hierarchical POS tag symbols

POS-tag<original form>	(allomorph)	[morphotactic adjacency symbols]
MCC<가공>	(가공)	[유>D하>H하>D되>]
MCK<거름>	(거름)	[유>D하>]
DI거라<건너가>	(건너가)	[규>축약>]
DIㄷ<알아듣>	(알아듣)	[규>]
DIㄷ<알아듣>	(알아들)	[불>어>]
DIㅅ<흐리젓>	(흐리저)	[불>어>]
DIㅅ<흐리젓>	(흐리젓)	[규>]
HIㄹ<가늘>	(가느)	[불>]
HIㄹ<가늘>	(가늘)	[규>어>]

Figure 2: Morpheme dictionary

pheme. To model the morpheme's connectability to one another, besides the morpheme dictionary, the separate morpheme-connectivity table encodes all the connectable pairs of morpheme groups using the morpheme's tag and morphotactic adjacency symbol patterns. After an input eojeol is segmented by trie indexed dictionary search, the morphological analysis checks if each segmentation is grammatically connectable by looking into the morpheme-connectivity table.

For unknown morpheme guessing, we develop a general unknown morpheme estimation method for number-free and position-free unknown morpheme handling. Using a morpheme pattern dictionary, we can look up unknown morphemes in the dictionary exactly same way as we do the registered morphemes. And when morphemes are checked if they are connectable, we can use the information of the adjacent morphemes in the same eojeol. The basic idea of the morpheme-pattern dictionary is to collect all the possible general lexical patterns of Korean morphemes and encode each lexical syllable pattern with all the candidate POS tags. So we can assign initial POS tags to each unknown morpheme by only matching the syllable patterns in the pattern dictionary. In this way, we don't need a special rule-based unknown morpheme handling module in our morphological analyzer, and all the possible POS tags for unknown morphemes can be assigned just like the registered morphemes. This method can guess the POS of each and every unknown morpheme, if more than one unknown morphemes are in an eojeol, regardless of their positions since the morpheme segmentation is applied to both the unknown morphemes and the registered morphemes dur-

ing the trie indexed dictionary search.

3.1 Morpheme pattern dictionary

The morpheme pattern dictionary covers all necessary syllable patterns for unknown morphemes including common nouns, proper-nouns, adnominals, adverbs, regular and irregular verbs, regular and irregular adjectives, and special symbols for foreign words. The lexical patterns for morphemes are collected from the previous studies (Kang, 1993) where the constraints of Korean syllable patterns as to the morpheme connectabilities are well described.
Fig. 3 shows some example entries of the morpheme pattern dictionary, where 'Z', 'V', '*' are meta characters which indicate a consonant, a vowel, and any number of Korean characters respectively. For example, "고마워" (thanks), which is a morpheme and an eojeol at the same time, is matched "(ZV*워)" (shown in Fig. 3) in the morpheme pattern dictionary, and is recovered into the original morpheme form "고맙".

4 A hybrid tagging model

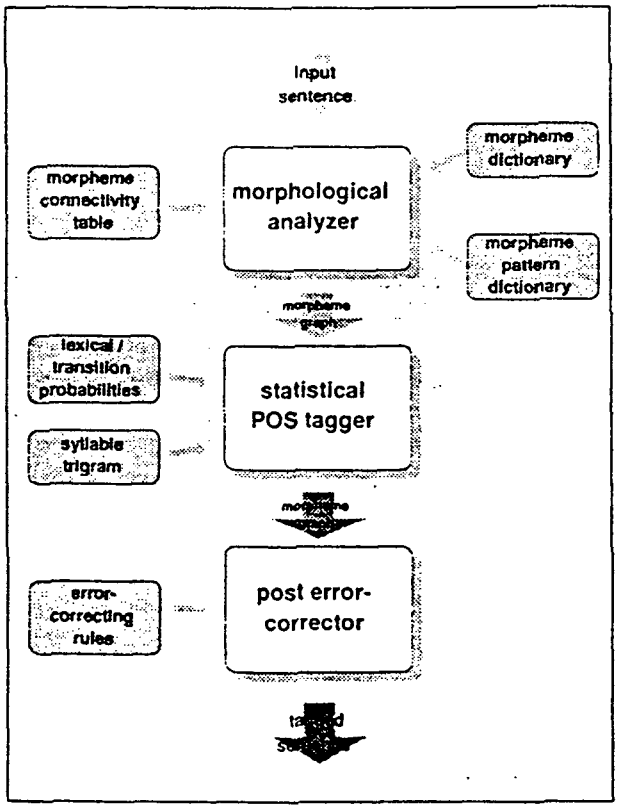

Figure 4: Statistical and rule-based hybrid architecture for Korean POS tagging.

Fig. 4 shows a proposed hybrid architecture for Korean POS tagging with generalized unknown-morpheme guessing. There are three major components: the morphological analyzer with unknown-morpheme handler, the statistical tagger, and the rule-based error corrector. The morphological analyzer segments the morphemes out of eojeols in a sentence and reconstructs the original morphemes from spelling changes from irregular conjugations. It also assigns all possible POS tags to each morpheme by consulting a morpheme dictionary. The unknown-morpheme handler integrated into the morphological analyzer assigns the POS's of the morphemes which are not registered in the dictionary.
The statistical tagger runs the Viterbi algorithm (Forney, 1973) on the morpheme graph for searching the optimal tag sequence for POS disambiguation. For remeding the defects of a statistical tagger, we introduce a post error-correction mechanism. The error-corrector is a rule-based transformer (Brill, 1992), and it corrects the mis-tagged morphemes by considering the lexical patterns and the necessary contextual information.

4.1 Statistical POS tagger

Statistical tagging model has the morpheme graph as input and selects the best morpheme and POS tag sequence[4] for sentences represented in the graph. The morpheme-graph is a compact way of representing multiple morpheme sequences for a sentence. We put each morpheme with the tag as a node and the morpheme connectivity as a link.
Our statistical tagging model is adjusted from standard bi-grams using the Viterbi-search (Cutting et al., 1992) plus on-the-fly extra computing of lexical probabilities for unknown morphemes. The equation of statistical tagging model used is a modified bi-gram model with left to right search:

$$T^* = argmax_T \prod_{i=1}^{n} \alpha Pr(t_i|t_{i-1}) \beta \frac{Pr(t_i|m_i)}{Pr(t_i)} \quad (1)$$

[4] A Korean eojeol can be segmented into many different ways, so selecting the best morpheme segmentation sequence is as important as selecting the best POS sequence in Korean POS tagging.

POS-tag<original form>	(allomorph)	[morphotactic adjacency symbols]
HIㄹ<ZV*갈>	(ZV*갈)	[규>어>]
HIㄹ<ZV*갈>	(ZV*가)	[불>]
HIㅂ<ZV*ZVㅂ>	(ZV*우)	[불>]
HIㅂ<ZV*ZVㅂ>	(ZV*워)	[축약>]
HIㅂ<ZV*ZVㅂ>	(ZV*와)	[축약>]
DIㅅ<ZV*젓>	(ZV*젓)	[규>]
DIㅅ<ZV*젓>	(ZV*저)	[불>어>]
DIㄷ<ZV*듣>	(ZV*듣)	[규>]
DIㄷ<ZV*듣>	(ZV*들)	[불>어>]

Figure 3: Morpheme pattern dictionary

where T^* is an optimal tag sequence that maximizes the forward Viterbi scores. $Pr(t_i|t_{i-1})$ is a bi-gram tag transition probability and $\frac{Pr(t_i|m_i)}{Pr(t_i)}$ is a modified morpheme lexical probability. This equation is finally selected from the extensive experiments using the following six different equations:

$$T^* = argmax_T \prod_{i=1}^{n} Pr(t_i|t_{i-1})Pr(m_i|t_i) \quad (2)$$

$$T^* = argmax_T \prod_{i=1}^{n} \alpha Pr(t_i|t_{i-1})\beta Pr(m_i|t_i) \quad (3)$$

$$T^* = argmax_T \prod_{i=1}^{n} Pr(t_i|t_{i-1})Pr(t_i|m_i) \quad (4)$$

$$T^* = argmax_T \prod_{i=1}^{n} \alpha Pr(t_i|t_{i-1})\beta Pr(t_i|m_i) \quad (5)$$

$$T^* = argmax_T \prod_{i=1}^{n} Pr(t_i|t_{i-1})\frac{Pr(t_i|m_i)}{Pr(t_i)} \quad (6)$$

$$T^* = argmax_T \prod_{i=1}^{n} \alpha Pr(t_i|t_{i-1})\beta\frac{Pr(t_i|m_i)}{Pr(t_i)} \quad (7)$$

In the experiments, we used 10204 morpheme training corpus from "Kemong Encyclopedia [5]". Table 1 shows the tagging performance of each equation.

Training of the statistical tagging model requires parameter estimation process for two parameters, that is, morpheme lexical probabilities and bi-gram tag transition probabilities. Several studies show that using as much as tagged corpora for training gives much better performance than unsupervised training using Baum-Welch algorithm (Merialdo, 1994). So we decided to use supervised training using tagged corpora with relative frequency counts. The three necessary probabilities can be estimated as follows:

$$Pr(t_i|m_i) \approx f(t_i|m_i) = \frac{N(m_i, t_i)}{N(m_i)} \quad (8)$$

$$Pr(t_i) \approx f(t_i) = \frac{N(t_i)}{\sum_{n=1}^{41} N(t_n)} \quad (9)$$

$$Pr(t_i|t_{i-1}) \approx f(t_i|t_{i-1}) = \frac{N(t_{i-1}, t_i)}{N(t_{i-1})} \quad (10)$$

where $N(m_i, t_i)$ indicates the total number of occurrences of morpheme m_i together with specific tag t_i, while $N(m_i)$ shows the total number of occurrences of morpheme m_i in the tagged training corpus. The $N(t_{i-1}, t_i)$ and $N(t_{i-1})$ can be interpreted similarly for two consecutive tags t_{i-1} and t_i.

4.2 Lexical probability estimation for unknown morpheme guessing

The lexical probabilities for unknown morphemes cannot be pre-calculated using the equation (8), so a special method should be applied. We suggest to use syllable tri-grams since Korean syllables can duly play important roles as restricting units for guessing POS of a morpheme. So the lexical probability $\frac{Pr(t_i|m_i)}{Pr(t_i)}$ for unknown morphemes can be estimated using the frequency of syllable tri-gram products according to the following formula:

$$m = e_1 e_2 ... e_n \quad (11)$$

[5] provided from ETRI

	equation 2	equation 3	equation 4	equation 5	equation 6	equation 7(equation 1)
eojeol	86.80	90.48	89.40	89.62	91.73	92.48
morpheme	91.32	94.93	94.40	94.48	95.77	96.12

Table 1: Tagging performance of each equation. The α and β are weights, and we set $\alpha = 0.4$ and $\beta = 0.6$. The eojeol shows eojeol-unit tagging correctness while morpheme shows morpheme-unit correctness.

$$\frac{Pr(t|m)}{Pr(t)} \approx Pr_t(e_1|\#,\#)Pr_t(e_2|\#,e_1) \prod_{i=3}^{n} Pr_t(e_i|e_{i-2},e_{i-1}) Pr(\#|e_{n-1},e_n) \quad (12)$$

$$Pr_t(e_i|e_{i-2},e_{i-1}) \approx f_t(e_i|e_{i-2},e_{i-1}) + f_t(e_i|e_{i-1}) + f_t(e_i) \quad (13)$$

where 'm' is a morpheme, 'e' is a syllable, 't' is a POS tag, '#' is a morpheme boundary symbol, and $f_t(e_i|e_{i-2},e_{i-1})$ is a frequency data for tag 't' with cooccurrence syllables e_{i-2}, e_{i-1}, e_i. A tri-gram probabilities are smoothed by equation (13) to cope with the sparse-data problem. For example, "박종만" is a name of a person, so is an unknown morpheme. The lexical probability of "박종만" as tag MPN is estimated using the formula:

$$\frac{Pr(MPN|\text{박종만})}{Pr(MPN)} \approx Pr_{MPN}(\text{박}|\#,\#) \times Pr_{MPN}(\text{종}|\#,\text{박}) \times Pr_{MPN}(\text{만}|\text{박},\text{종}) \times Pr_{MPN}(\#|\text{종},\text{만})$$

All tri-grams for Korean syllables were precalculated and stored in the table, and are applied with the candidate tags during the unknown morpheme POS guessing and smoothing.

5 A posteriori error correction rules

The statistical morpheme tagging covers only the limited range of contextual information. Moreover, it cannot refer to the lexical patterns as a context for POS disambiguation. As mentioned before, Korean eojeol has very complex morphological structure so it is necessary to look at the functional morphemes selectively to get the grammatical relations between eojeols. For these reasons, we designed error-correcting rules for eojeols to compensate estimation and modeling errors of the statistical morpheme tagging. However, designing the error-correction rules with knowledge engineering is tedious and error-prone. Instead, we adopted Brill's approach (Brill, 1992) to automatically learn the error-correcting rules from small amount of tagged corpus. Fortunately, Brill showed that we don't need a large amount of tagged corpus to extract the symbolic tagging rules compared with the case in the statistical tagging. Table 2 shows some rule schemata we used to extract the error-correcting rules, where a rule schema designates the context of rule applications, i.e., the morpheme position and the lexical/tag decision in the context eojeol.

The rules which can be automatically learned using table 2's schemata are in the form of table 3, where [*current eojeol or morpheme*] consists of morpheme (with current tag) sequence in the eojeol, and [*corrected eojeol or morpheme*] consists of morpheme (with corrected tag) sequence in the same eojeol. For example, the rule [먹$(Chinese\ ink)/MC$ + 은$/jS][$N1FT, MC$] → [먹$(to\ eat)/DR$ + 은$/eCNMG$] says that the current eojeol was statistically tagged as common-noun (MC) plus auxiliary particle (jS), but when the next first eojeol's (N1) first position morpheme tag (FT) is another common-noun (MC), the eojeol should be tagged as regular verb (DR) plus adnominal ending (eCNMG). This statistical error is caused from the ambiguity of the morpheme "먹" which has two meanings as "Chinese ink" (noun) and "to eat" (verb). Since the morpheme segmentation is very difficult in Korean, many of the tagging errors also come from the morpheme segmentation errors. Our error-correcting rules can cope with these morpheme

rule schema	description
N1FT	next first eojeol (N1) first morpheme's tag (FT)
P1LT	previous first eojeol (P1) last morpheme's tag (LT)
N2FT	next second eojeol (N2) first morpheme's tag (FT)
N3FT	next third eojeol (N3) first morpheme's tag (FT)
P1LM	previous first eojeol (P1) last morpheme's lexical form (LM)
P1FM	previous first eojeol (P1) first morpheme's lexical form (FM)
N1FM	next first eojeol (N1) first morpheme's lexical form (FM)

Table 2: Some rule schemata to extract the error-correcting rules automatically from the tagged corpus. POSTAG has about 24 rule schemata in this form.

[current eojeol or morpheme] [rule schemata, referenced morpheme or tag] → [corrected eojeol or morpheme]

Table 3: Error correction rule format

segmentation errors by correcting the errors in the whole eojeol together. For example, the following rule can correct morpheme segmentation errors: [줄/MC + 이고/jO][$P1LM$, 을] → [줄이/DR + 고/eCC]. This rule says that the eojeol "줄이고" is usually segmented as common-noun "줄" (meaning string or rope) plus other-particle "이고", but when the morpheme "을" appears before the eojeol, it should be segmented as regular-verb "줄이" (meaning shrink) plus conjunctive-ending "고". This kind of segmentation-error correction can greatly enhance the tagging performance in Korean. The rules are automatically learned by comparing the correctly tagged corpus with the outputs of the statistical tagger. The training is leveraged (Brill, 1992) so the error-correcting rules are gradually learned as the statistical tagged texts are corrected by the rules learned so far.

6 Experiment results

For morphological analysis and POS tagging experiments, we used 130000 morpheme-balanced training corpus for statistical parameter estimation and 50000 morpheme corpus for learning the post error-correction rules. These training corpora were collected from various sources such as internet documents, encyclopedia, newspapers, and school textbooks.

For the test set, we carefully selected three different document sets aiming for a broad coverage. The document set 1 (25299 morphemes; 1338 sentences) is collected from "Kemong encyclopedia [6]", hotel reservation dialog corpus[7] and internet document, and contains 10% of unknown morphemes. The documents set 2 (15250 morphemes; 574 sentences) is solely collected from various internet documents from assorted domains such as broadcasting scripts and newspapers, and has about 8.5% of unknown morphemes. The document set 3 (20919 morphemes; 555 sentence) is from Korean standard document collection set called KTSET 2.0[8] and contains academic articles and electronic newspapers. This document set contains about 14% unknown morphemes (mainly technical jargons).

Table 4 shows our tagging performance for these three document sets. This experiment shows efficiency of our unknown morpheme handling and guessing techniques since we can confirm the sharp performance drops between tagger-a and tagger-b. The post error correction rules are also proved to be effective by the performance drops between the full tagger and tagger-a, but the drop rates are mild due to the performance saturation at tagger-a, which means that our statistical tagging alone already achieves state-of-the-art performance for Korean morpheme tagging.

[6] from ETRI

[7] from Sogang University, Seoul, Korea

[8] from KT(Korea Telecom)

document set	full tagger	tagger-a	tagger-b	tagger-c
set 1	97.2	96.4	89.5	87.1
set 2	96.9	96.0	92.8	89.0
set 3	97.4	96.7	88.7	84.8
total	97.2	96.4	90.3	87.0

Table 4: Tagging and unknown morpheme guessing performance (all in %). Experiments are performed on three different document sets as explained in the text. The full tagger designates our POS tagger with all the morphological processing capabilities. The tagger-a is a version without employing post error-correction rules. The tagger-b is a more degraded version which does not utilize our unknown morpheme guessing capability but treats all unknown morphemes as nouns. The tagger-c is an even more deteriorated version which rejects all unknown morphemes as tagging failures. The performance drops as we degrade the version from the full tagger.

7 Conclusion and future works

This paper presents a pattern-dictionary based unknown-morpheme guessing method for a statistical/rule-based hybrid tagging system which itself exhibits many novel ideas of POS tagging such as experiment-based new statistical model for Korean, rule based error correction and hierarchically expandable tag sets. The system POSTAG was developed to test these novel ideas especially for agglutinative languages such as Korean. Japanese is also similar to Korean in linguistic characteristics and will be a good target of these ideas. POSTAG integrates morphological analysis with generalized unknown-morpheme handling so that unknown-morpheme can be processed in the same manner as registered morphemes using morpheme pattern dictionary. POSTAG adopted a hybrid approach by cascading statistical tagging to rule-based error-correction. Cascaded training was implemented to selectively learn statistical tagging error-correction rules by Brill style transformation approach. POSTAG also employs hierarchical tag sets that are flexible enough to expand/shrink according to the given applications. The hierarchical tag sets can be mapped to any other existing tag set as long as they are decently classified, and therefore can encoverage a corpus sharing in Korean tagging community. POSTAG is constantly being improved by expanding the morpheme dictionary, pattern-dictionary, and tagged corpus for statistical training and rule learning. Since generalized unknown-morpheme handing is integrated into the system, POSTAG is a good tagger for open domain applications such as internet indexing, filtering, and summarization, and we are now developing a web indexer using the POSTAG technology.

References

E. Brill. 1992. A simple rule-based part-of-speech tagger. In *Proceedings of the conference on applied natural language processing.*

E. Brill. 1994. Some advances in transformation-based part-of-speech tagging. In *Proceedings of the AAAI-94.*

D. Cutting, J. Kupiec, J. Pedersen, and P. Sibun. 1992. A practical part-of-speech tagger. In *Proceedings of the conference on applied natural language processing.*

G. Forney. 1973. The Viterbi algorithm. *Proc. of the IEEE*, 61:268–278.

S. S. Kang. 1993. *Korean morphological analysis using syllable information and multiple-word units.* Ph.D. thesis, Department of Computer Engineering, Seoul National Univeristy. (written in Korean).

J. Kupiec. 1992. Robust part-of-speech tagging using a hidden markov model. *Computer speech and language*, 6:225–242.

B. Merialdo. 1994. Tagging English text with a probabilistic model. *Computational linguistics*, 20(2):155–171.

R. Sproat. 1992. *Morphology and computation.* The MIT Press, Cambridge, MA.

A. Voutilainen. 1995. A syntax-based part-of-speech analyzer. In *Proceedings of the seventh conference of the European chapter of the association for computational linguistics (EACL-95)*, pages 157–164.

R. Weischedel, M. Meteer, R. Schwartz, L. Ramshaw, and J. Palmucci. 1993. Coping with ambiguity and unknown words through probabilistic model. *Computational linguistics*, 19(2):359–382.