

RULE-BASED MORPHOLOGICAL DISAMBIGUATION USING STATISTICS IN AN AGGLUTINATIVE LANGUAGE †

Aesun Yoon, Hyuk-Chul Kwon*

Dept. of French, Dept. of Computer Science, *
Pusan National University, Rep. of Korea
{asyoon, hckwon}@hyowon.pusan.ac.kr

Abstract: Current Korean morphological disambiguation systems adopt mainly statistical methods. Some of them use rules in the postprocess. In our approach, the morphological analyzer reduces the number of candidate morpheme strings by using adjacency conditions during the analysis of a word into morpheme strings. Disambiguation then depends on rules and statistics successively. The accuracy rate of this rule-based Korean language tagging system is about 97.1%. It yields a better result than morphological disambiguation systems.

Keywords: Korean, Agglutinative Language, Morphology, Disambiguation, Rule-Based Parsing, Statistics, Adjacency Conditions, Computational Linguistics

1. INTRODUCTION

The purpose of this paper is to show efficiency of rule-based approach, compared to statistical methods, in disambiguation of an agglutinative language. Current Korean morphological disambiguation systems adopt mainly statistical methods. Some of them use rules in the postprocess to filter out incompatible strings. (Lee, Seo and Oh, 1996; Lim, Kim and Rim, 1996; Lee and Lee, 1996) The statistical approach can, however, state only positive adjacency conditions on two morphemes. It fails to predict accurately their mutual exclusion that plays an important role in morphological disambiguation. (Kwon & Chae, 1991)

As an agglutinative language, Korean has its own types of morphological ambiguities. A large number of Korean ambiguities is related to segmentation into stems and endings, as

ISBN: 0 08 043 438X

shown in the examples below (1), while most of ambiguities in French or English are due to the categorization of a morpheme. (Chanod and Tapanainen, 1995) When a Korean word is analyzable into several different morpheme strings, it is not easy to decide which one is the most appropriate within the context. (Lim, Lee and Rim, 1993)

(1) *gan* (간)
(a) *ga* ("go/change/crush"[verb]) + *n* ([ending])
(b) *gan* ("liver"[noun])
(c) *ga* ("edge/saltness/room"[noun]) + *n* ([topic marker])
(Where the symbol {+} represents the morpheme boundary)

Moreover, Korean orthography spacing rules are optional in some cases,(Mee, 1994) or they are not strictly respected. The author decides whether or not to put spaces between two nouns in a noun compound. More than 10% of words in newspapers violate the spacing rules to save space. (Kwon, 1995) For example, the string *gan-ji-nwn* without a space has two interpretations, (2-a) and (2-b). A space should be put between *gan* and *ji*, when the morpheme *ga* is a verb as in (2-c). Since this spacing rule is frequently violated, a Korean morphological analyzer should be able to distinguish among three different interpretations of *gan-ji-nwn.*

(2) *gan-ji-nwn* (간지는)
(a) *ganji* ("writing paper/the sexagenary cycle"[noun]) + *nwn* ([topic marker])
(b) *ga* ("edge"[noun]) + *n-ji* ("whether (or not)"
[ending - contracted form of *in-ji*]) + *nwn* ([topic marker])
(c) *ga* ("go/change/crush" [verbe]) + *n* ([ending]) # *ji* ([bound noun])
+ *nwn* ([topic marker])
(where the symbols {-, #} represent the syllable boundary and the word boundary, respectively.)

Current morphological disambiguation systems, however, fail to predict (2-c), which will be an acceptable interpretation in most of contexts. In our approach, as shown in Fig. I, the morphological analyzer reduces the number of the candidate morpheme strings by using adjacency conditions during the analysis of a word into morpheme strings. The disambiguation then depends on rules and statistics applied successively. As for the rules, the partial parsing using finite state automata decides the compatibility of each morpheme pair: a negative value is assigned when two morphemes cannot co-occur, while a positive value is given if they are compatible. After applying all the rules related to the word, our system chooses only positively valued strings. When more than one strings still have the same value, a statistical process determines the context priority in the next step. If a given word fails to be disambiguated, or if the evaluation value is very low, our system involves the guessing routine. These 4 successive steps will be described in this paper.

ISBN: 0 08 043 438X

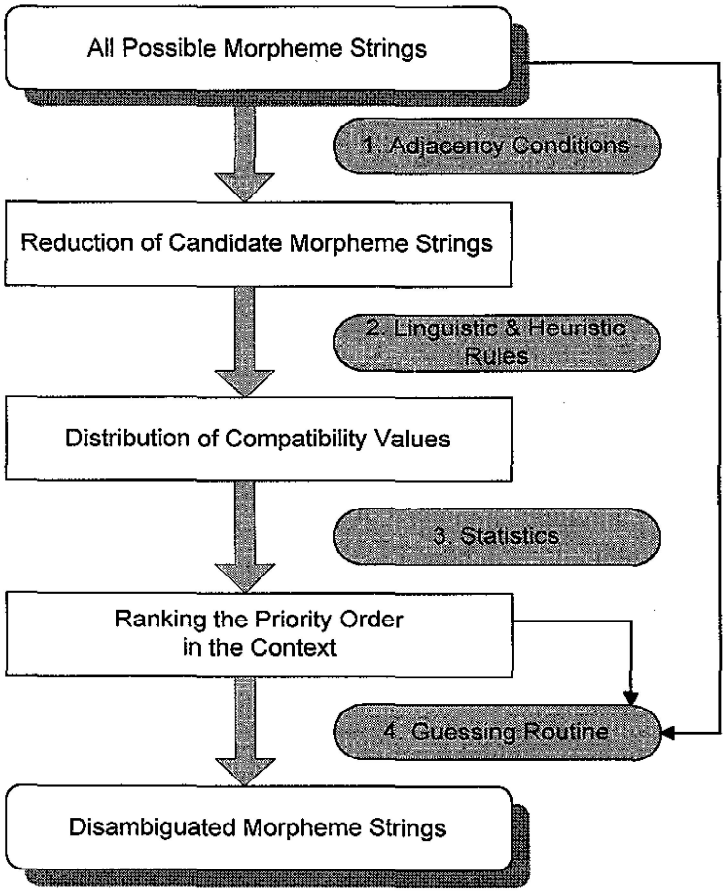

Fig. I. Disambiguation Process Using Rules and Statistics

2. REDUCTION OF CANDIDATE MORPHEME STRINGS

The efficiency and the speed of a morphological disambiguation system depends largely on the output of a morphological analyzer. When a Korean morphological analyzer segments a word into morpheme strings, one of the serious problems is over-analysis. To filter out incompatible morpheme strings, our morphological analyzer applies, before the disambiguation process, adjacency conditions which can determine the compatibility or the incompatibility of two morphemes.

The domain of adjacency conditions in our morphological analyzer is **intra-word**, i.e. two morphemes that constitute one (compound) word, occurring between two spaces. Adjacency conditions on morpheme pairs can be described either by constraints or by lists. Constraints determine compatibility if one constituent of the morpheme pair has a strong generative power, such as *'-wm'* or *'-in'* as in (3-a,b). If the distribution is restricted, all its morphemes-pairs are listed, as in (4-a,b).

(3) (a) [verb stem] + ***wm*** ([nominalization suffix]):
(ex) meok-wm (먹음 "to eat"), *al-wm* (알음 "to know")
(b) name of a country + ***in*** ("person"[noun suffix]):
(ex) pw-lang-sw-in (프랑스인 "French"),
han-guk-in (한국인 "Korean")

(4) (a) name of a government office + ***seo*** ("branch"[noun suffix]):

ISBN: 0 08 043 438X

(ex) kyeong-chal-seo (경찰서 "police station"),
so-bang-seo (소방서 "fire station")

(b) name of a government office + ***so*** (place[noun suffix]):
(ex) pha-chul-so(파출소, "police box"),
po-keon-so(보건소, "health center")

Adjacency conditions in our morphological analyzer focus mainly on nouns. (Church, 1993) They are formulated in terms of constraints on location within a compound word, in terms of constraints on one-syllable nouns, and constraints on noun suffixes.

2.1. Location of Nouns within Noun Compounds

'No-dong-ja-ga' shows a high potential for producing ambiguities, which can be resolved only by semantic analysis

(5) no-dong-ja-ga (노동자가)
(a) *no-dong* ("labor"[noun]) + *-ja* ("person"[noun]) + *ga* ([subject marker])
(b) **no-dong* ("labor"[noun]) + *ja-ga* ("private house"[noun])

If adjacency condition (6) on the location of the noun '*ja-ga* ("private house"[noun])' states that its distribution is restricted to the first place of a compound noun, *no-dong-ja-ga* will be interpreted only as (5-a), and (5-b) will be filtered out. Our morphological analyzer contains adjacency conditions on 1,700 such nouns, in which the location of noun within the compound is decisive.

(6) Condition on '*ja-ga* ("private house"[noun])' :
{It should be the first component of a compound word.}

2.2. One-Syllable Noun

The morpheme string '[one syllable noun] + [case marker or postposition]' is the cause of the most frequently occurred ambiguities. The example (7) '*su-lwl'* might be analyzed as (7-a) and (7-b).

(7) *su-lwl* (수를)
(a) *su* ("number"[noun]) + *lwl* ([object marker])
(b) **su* ([bound noun]) + *lwl* ([object marker])

Adjacency condition (8) requires the bound noun '*su'* to be followed by an intransitive verb. Thus *'su'* cannot co-occur with an object marker. Our system generates only (7-a) from '*su-lwl'* and filters out (7-b).

(8) Condition on the Bound Noun '*su'*:
{It should be followed by an intransitive verb '*iss-'* or '*eobs-'*.}

2.3. Noun Suffixes

Frequently occurring spelling errors must also be corrected by the morphological

ISBN: 0 08 043 438X

analyzer. According to Korean orthography, a space must be used between '*su*("number"[noun])' and the preceding noun. But a large number of examples in our corpus of data violates this rule. In the example (9), current Korean morphological analyzers might incorrectly generate only (9-a) from '*†no-dong-ja-su*' and fail to predict (9-b) which is correct.

(9) *†no-dong-ja-su* (노동자수)
(a) **no-dong* ("labor"[noun]) + *ja-su* ("embroidery/self-surrender"[noun])
(b) *no-dong*("labor"[noun]) + *ja* ("person"[noun suffix]) # *su* ("number"[noun])
(Where { † } represents the violation of spacing rules.)

Another type of our adjacency conditions (10) gives priority to the parts-of-speech containing noun suffixes such as '*-ja*', '*-ga*', etc. followed by '*su* ("number[noun])', even though the word is orthographically incorrect. Our system not only generates (9-b) from '*no-dong-ja-su*', but also assigns a positive value to it.

(10) Condition on '*–ja* ("person"[noun suffix])' or '*–ga* ("person"[noun suffix])'
{'*su* ("number"[noun])' has a priority in the interpretation,
if it is preceded by '*–ja*' or '*–ga*'}

Compared with the other disambiguation systems, our system, using adjacency conditions, can reduce the number of candidate morpheme strings. When our system analyses successfully more than 99% of the corpus, 33.2% of its output shows ambiguities. The average interpretation number per one ambiguous word is 2.75. That means the average number of candidate morpheme strings is only 1.58 per one word.

3. DISAMBIGUATION BY RULE-BASED APPROACH

When a word still remains ambiguous after having been processed by the morphological analyzer, our system depends on three different types of rules for disambiguation. The domain of those rules is **inter-word.** i.e. these rules operate within the unit 'word'. They predict the compatibility conditions of a morpheme with its preceding and/or following word(s). The rules focus on a governing morpheme and parse its left and right context. Even though the window size of the context is determined by the linguistic constraints on that morpheme, users can reduce the window size to speed up the disambiguation process.

3.1. Rules on Highly Ambiguous Morphemes

The first type of rules concerns 63 specific morphemes (or parts-of-speech), such as '*dae-ha-da* ("be over against"[verb], and its conjugated forms)', '*han* ("one"[adjective]/ "done"[verb]/ "heart-burning"[noun])' and '*su* ("number"[noun]/ [bound noun])', since we found that more than 27% of all the ambiguous words in our corpus were related to these 63 morphemes in some way. After describing the rules to disambiguate these morphemes and/or their related morphemes, we test these rules on a large corpus of data and refine them. These rules assign either a positive or a negative value to each morpheme combination, and its following and/or preceding morphemes.

3.2. Syntactic Constraints

ISBN: 0 08 043 438X

The second type of rules is related to 15 syntactic constraints on morphemes, as follows: (Nam and Go, 1986)

(11) (a) Bound nouns must follow a word with an adjectival ending;
(b) An adjectival ending should be followed by a noun, or by another word with an adjectival ending;
(c) An intransitive verb, or an adjective, may not follow a noun with an objective marker, if they also have an adjectival ending;
(d) Declarative endings should precede a quotation mark;
(e) etc.

Each of these rules has a different constraint power, that affects the assignment of the positive value to the morpheme when the rule is satisfied. The rule (11-a) is so strong that the system assigns a high positive value whereas the rule (11-c), less strong than the rule (11-a), gives a relatively low value. A negative value is assigned whenever any of these rules is not satisfied.

3.3. Heuristic Rules

The third type of rules uses the collocation of morphemes. The stem '*ssw-*' can belong to two different categories: a verb ("write/use") and an adjective ("bitter"). When its subject is a medicine, a plant, a food, etc., it must be an adjective. But if its subject is a human being, it must be a verb. We are now in the process of formulating similar heuristic rules for the the purpose of disambiguation, focussing on 200,000 high frequency words. These 200,000 words form 87% of our corpus of 11 million words.

3.4. Disambiguation Process

The following shows the disambiguation process using our rule-based system, explained above, of ambiguous strings '*al-su-nwn iss-ta*' and '*al su-nwn morw-ta*'. After being processed by the morphological analyzer, two strings are segmented into 6 morphemes as shown in Fig. II, of which '*al*' and '*su*' are ambiguous. But each string must have only one interpretation instead of two possible ones.

In the first step of disambiguation, as shown in Fig. III, the morpheme '*al*', as a noun, does not give any constraint on the following morphemes, while the rule (11-b) tells that '*al+l*', as an adjectival form, should be followed by a noun, or a word with an adjectival ending.

According to rule (11-a), '*su2*', as a bound noun, requires that the preceding word must be an adjectival form as shown in Fig. IV. Thus '*al*' has a negative value with respect to '*su2 + nwn*', while '*al + l*' has a positive value. In the second step, the system chooses only '*al-l su2-nwn*', if no other words follow.

ISBN: 0 08 043 438X

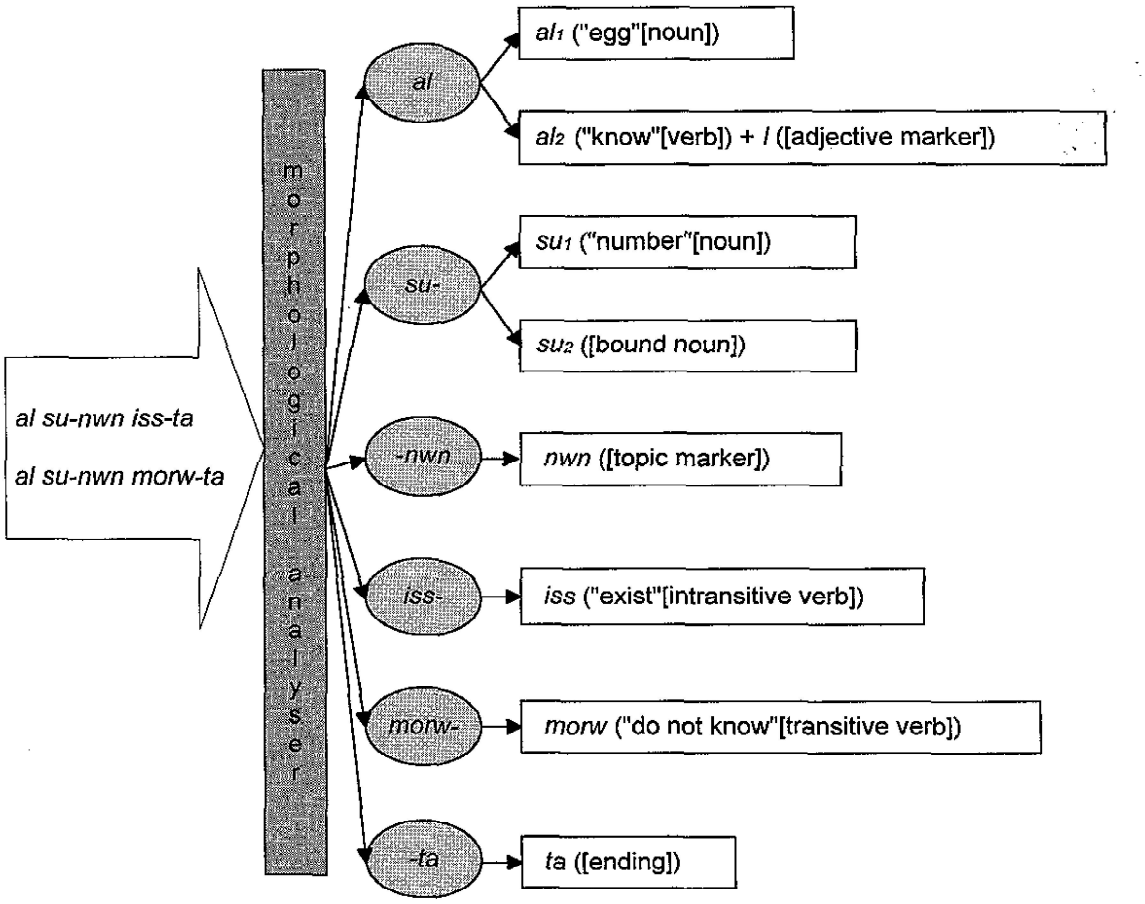

Fig. II : Segmentation into Morphemes by Morphological Analyzer

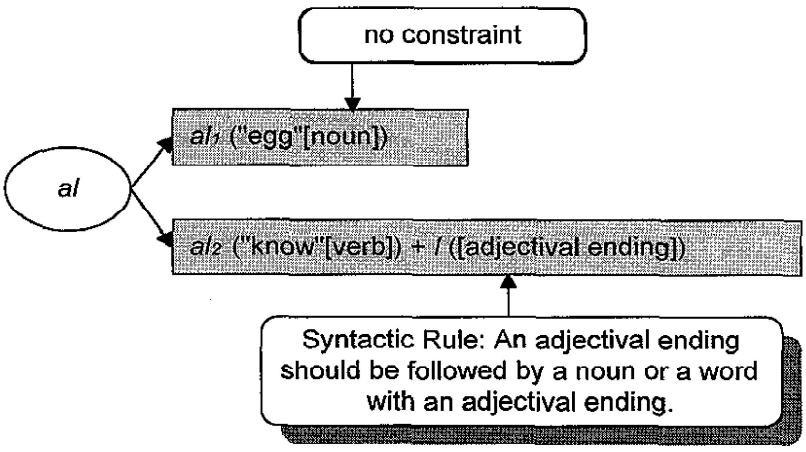

Fig. III : Syntactic Rule on Adjectival endings

ISBN: 0 08 043 438X

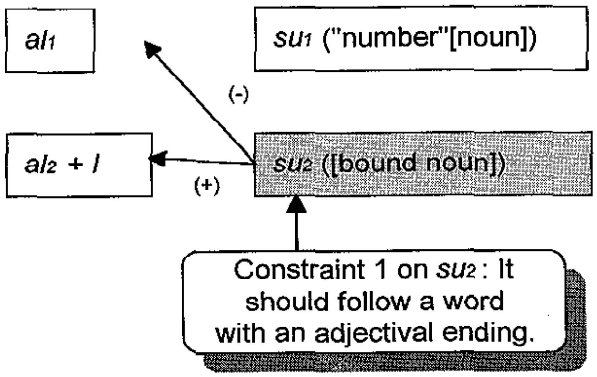

Fig. IV : Constraint 1 on Bound Noun '*su2*'

Suppose that '*al-l su-nwn*' is followed by other words. Another constraint concerning '*su2*' states that the following verb must be '*iss* ("exist"[intransitive verb])' or '*eobs* ("do not exist"[intransitive verb])'.

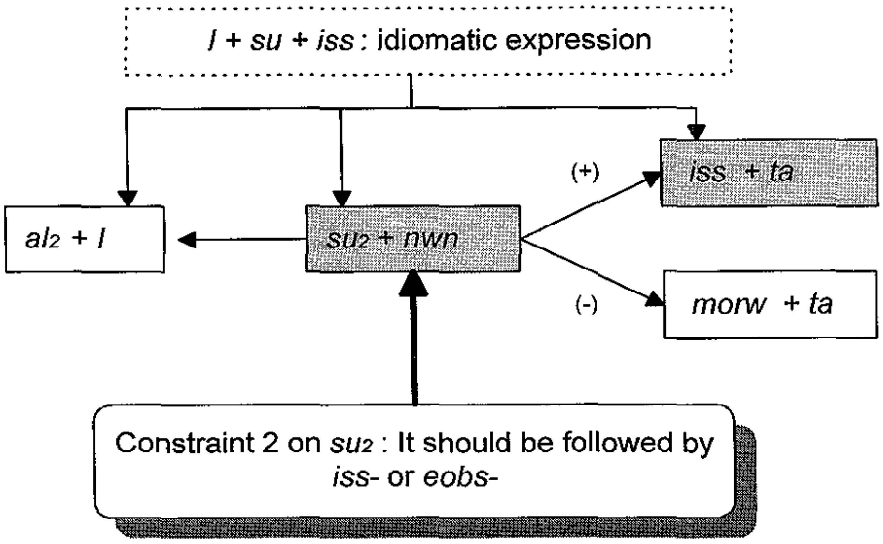

Fig. V : Constraint 2 on Bound Noun '*su2*'

Since '*-l su2 iss* (can)', defined as an idiomatic expression, assigns a positive value to the link between '*su2-nwn*' and '*iss-ta*', the highest priority will be given to the string '*al-l su2-nwn iss-ta* ("can know")'. The same rule assigns a negative value to the pair '*su2-nwn*' and '*mo-lw-ji*'. '*Al su1-nwn mo-lw-ji* ("do not know the number of eggs")' is the unique output, because our system filters out all negatively valued morpheme strings.

4. SELECTION OF THE MOST FEASIBLE MORPHEME STRING BY STATISTICS

When the system fails to disambiguate a word by applying rules on collocation of morphemes, the most acceptable morpheme string is selected according to properties of the morphemes constituting one word. Although both the morphological analyzer and the selection routine use only the **intra-word** information, this selection routine is different from the morphological analyzer. While the morphological analyzer uses the constraints in order to remove incorrect morpheme strings, the selection process depends on the property

ISBN: 0 08 043 438X

information to assign priority order for all the remaining morpheme strings.

The selection routine employs, as intra-word information, the frequency of a morpheme within the corpus, individual morpheme category pattern within strings, and information on special morphemes that can affect the morpheme string selection.

4.1. Frequency of Morpheme

At the beginning of our research, we expected that the following frequency-based evaluation function could effectively select the most acceptable morpheme string.

(12) $F(M_1M_2\ldots M_n) = \prod_{i=1}^{n} P(M_i)$

(M_i : i-th morpheme in a morpheme string
$P(M_i)$: the frequency of M_i in the corpus)

However, this function cannot be used as a general heuristic method for the following two reasons.

First, the high frequency does not always predict the high priority. Some morphemes such as '*-l* ([object marker])' and '*-n* ([topic marker])', which are contracted forms of '*wl*' and '*wn*' respectively, have high frequency. For example, (13-a', b', c', d'), the contracted forms of (13-a, b, c, d), occur frequently in the corpus.

(13) (a) *i-geos* ("this one" [pronoun]) + *wl* ([object marker])
(a') *i-geo-l*
(b) *jeo-geos* ("that one" [pronoun]) + *wl* ([object marker])
(b') *jeo-geo-l*
(c) *i-geos* ("this one" [pronoun]) + *wn* ([topic marker])
(c') *i-geo-n*
(d) *jeo-geos* ("that one" [pronoun]) + *wn* ([topic marker])
(d') *jeo-geo-n*

The high frequency of these morphemes was unexpected. Since the distribution of contracted forms '*-l* ' or '*-n*' is extremely limited, most of them are unambiguous, even though their frequency in the corpus is high. With the function as described in (12), the system is more likely to select (14-a) than (14-b), which is the correct interpretation. But in ambiguous cases, such as (14), the morpheme string without '*-n*' and '*-l*' has a high selection priority.

(14) *yeo-bun* (여분)
(a) *yeo-bu* ("whether" [noun]) + n ([topic marker – contracted form])
(b) *yeo-bun* ("superfluity" [noun])

Second, the evaluation value of the morpheme strings containing a morpheme that does not appear in our corpus becomes 0. For example, '*so-jil*' can be analyzed as (15-a) and (15-b). When the corpus does not contain (15-b), a morpheme string with '*-l*' is selected, even if the distribution of (15-a) is extremely restricted. The morpheme string (15-a) is correct only when '*ha-da* ("do" [verb])' follows it. This case is covered by the collocation rules.

ISBN: 0 08 043 438X

(15) *so-jil (소질)*
(a) *so-ji* ("possession" [noun]) + *l* ([object marker-contracted form])
(b) *so-jil* ("talent" [noun])
(16) *gam-gag-gi-do (감각기도)*
(a) *gam-gag* ("sense" [noun]) + *gi-do* ("pray"[noun])
(b) *gam-gag-gi* ("sensor" [noun]) + *do* ("also"[postposition])

As for '*gam-gag-gi-do*', which has two interpretations, only (16-a) is selected by the evaluation function, if '*gam-gag-gi* (sensor)' is absent from the corpus of 11 million words. Although (16-a) is not semantically correct, our morphological analyzer might not remove this interpretation since it does not use semantic information. The degeneration due to absence of some morphemes from the corpus should be considered in order to achieve a successful disambiguation rate of more than 97%.

4.2. Preference Patterns Using Category Information

In our experiments, using category information for morpheme string patterns produces better results. Our selection routine prefers '[noun] + [postposition]' to '[noun] + [noun]' or '[noun]', if the postposition is not a contracted form (i.e. neither '*-l*' nor '*-n*'), and if the noun does not consist of one syllable only.

The one-syllable noun is handled in our system in a special way, since it presents great difficulties in morphological disambiguation of Korean. For example, '*tho-wi* ("discussion"[noun])' is preferred rather than '*tho* ("soil"[noun]) + *wi* ("of" [postposition])'.

Preference patterns are based on the results of statistical analysis of the patterns occurring in the corpus. Unambiguous words provide the statistical information. Words like '*gam-gag-gi-do*' and '*so-jil*' can be successfully disambiguated by preference patterns. We use the evaluation function however when the patterns are similar. For example, since '[verb] + [ending]' and '[adjective] + [ending]' are regarded as similar patterns, '*sseo-seo*' is analyzed as both (17-a) and (17-b). Since '*ssw* ("write"[verb])' appears more frequently than '*ssw*("bitter"[adjective])', selection routine gives priority to (17-a), if no other rule disambiguated it before.

(17) *sseo-seo* (써서)
(a) *ssw*("write" [verb]) + *eo-seo*([ending])
(b) *ssw*("bitter" [adjective]) + *eo-seo*([ending])

The evaluation function also solves ambiguities caused by stemming of a compound noun or a compound predicate. For disambiguation of the stemming, frequencies of the morphemes in morpheme strings are multiplied. As a result, the word composed of fewer morphemes is preferred if all morpheme frequencies are identical. The preference for fewer morphemes is a generalized rule that can disambiguate Korean or Japanese compound words.

4.3. Special Morpheme

If a given word fails to be disambiguated, or if the evaluation value is very low, our system involves the guessing routine. It guesses meaning of an unknown word (or parts of speech) by removing plausible postpositions or endings attached to the word. If the system

ISBN: 0 08 043 438X

guesses more than one different strings for the unknown word, it selects the most appropriate one on the basis of frequency of their postpositions or endings.

The accuracy of the guessing routine for the unknown words (or parts of speech) is about 98% and the successful disambiguation rate of our system is about 97.1% excluding unknown words.

V. CONCLUSION

The approach described in this paper is different from the approach that is currently used for Korean morphological disambiguation: the rules are applied first and the statistical method is supplementary. Although the rule-based approach is difficult to implement, we feel confident that the accuracy rate can be improved, if we provide much more information to the system.

We also assume that preparing a high-quality tagged corpus for Korean is a much more difficult task than formulating linguistic rules. By using only linguistic and heuristic rules, we can achieve accuracy rate of about 95.3%. This accuracy is very high compared with the statistical method-based systems. Until now, no disambiguation system depending solely on statistical methods has exceeded the accuracy rate of 93%. And the demon programming promises high processing speed of our system. That is, the dictionary information of the morphemes (or parts of speech) incorporates the rules that apply the ambiguities related to them. Furthermore our system does not use any domain specific rules, it is more robust than the statistical methods.

Our system using the combination of rules and statistical data yield a better result than other Korean morphological disambiguation systems: the accuracy rate is about 97.1% for middle school and high school textbook data.

REFERENCES

Chanod, Jean-Pierre and Pasi Tapanainen, 1995. *Tagging French - comparing a statistical and a constraint-based method. cmp-lg/9503003*, 1995.

Church, Kenneth W., 1993. A Stochastic Parts Program and Noun Phrase Parser for Unrestricted Text. *Proc. of the Second Conference on Applied Natural Language Processing*, 136-143.

Kwon, Hyuk-Chul and Young-Suk Chae. 1991. A Dictionary-based Morphological Analysis, *Proc. of NLPRS '91*, 141-147.

Kwon, Hyuk-Chul. 1995. *Development of Algorithms for the system to support Writing and Revising Texts*. System Engineering Research Institute.

Lee, Geun-Bae and Jong-Hyeok Lee. 1996. A Robust Statistical Part-of-Speech Tagging System for Korean Texts. *Proc. of the Second Korea-China Joint Symposium on Oriental Language Computation '96*, 125-131.

Lee, Sang-Ho, Jung-Yun Seo and Yung-Hwan Oh. 1996. A Robust Statistical Part-of-Speech Tagging System for Korean Texts. *Proc. of the Second Korea-China Joint Symposium on Oriental Language Computing '96, 111-118.*

Lim, Heui-Seok, Ho Lee and Hae-Chang Rim. 1993. A Method of Analyzing Word Ambiguity in Korean Morphological Analysis. *Proc. of the Korean Information Science Society,* **Vol. 20 No. 1**, *703-776*

Lim, Heui-Seok, Jin-Dong Kim and Hae-Chang Rim. 1996. *Proc. of the Second Korea-China Joint Symposium on Oriental Language Computing '96, 119-124.*

ISBN: 0 08 043 438X

Nam, Ki-Sim and Young-Geun Go. 1986. *Standard Korean Grammar.* Top Publisher:Seoul.
Seung-Woo Mee. 1994. *New Korean Orthography.* Emunkak:Seoul.

† This paper has been supported in part by Korea Science & Engineering Foundation (project number 96-2-11-02-01-3) and by the Research Institute of Computer & Information-Communication of Pusan National University.

ISBN: 0 08 043 438X