Language, Information and Computation(PACLIC 11), 1996, 237-246

Rule-based Approach to Korean Morphological Disambiguation Supported by Statistical Method

Min-Jung Kim* Hyuk-Chul Kwon* Ae-Sun Yoon**
Dept. of Computer Science*, Dept. of French**
Pusan National University, Korea
{mjkim, hckwon, asyoon}@bandi.cs.pusan.ac.kr

ABSTRACT

Korean as an agglutinative language shows its proper types of difficulties in morphological disambiguation, since a large number of its ambiguities comes from the stemming while most of ambiguities in French or English are related to the categorization of a morpheme.

The current Korean morphological disambiguation systems adopt mainly statistical methods and some of them use rules in the postprocess. In our approach, the morphological analyzer reduces the number of the candidate morpheme strings using adjacency conditions when it analyses a word into morpheme strings. And then the disambiguation depends on rules and statistics successively. As for the rules, the partial parsing using finite state automata decides the compatibility of each pair of words: a negative value is assigned if a word can not co-occur with another word, while a positive value is given if they are compatible. After applying all the rules related to the word, our system chooses only the positively valued strings. When more than two strings still have same value, the priority in the context is decided by the statistics in the next stage. The accuracy of our approach as Korean tagging system is about 97.1% and it may yeild a better result than the Korean morphological disambiguation systems.

I. Introduction

Compared to French or English, Korean as an agglutinative language shows its proper types of difficulties in morphological disambiguation, since a large number of its ambiguities comes from the stemming while most of ambiguities in French or English are related to the categorization of a morpheme[1]. In the case that a Korean word can be analyzed into several different morpheme strings[2], it is not easy to decide which one is the most compatible with the context. And what is worse, spacing rules of the Korean orthography are optional in some cases or they are easily violated in the other cases: it depends

on authors to give spaces or not between two nouns in noun compounds; more than 10% of words in newspapers violate the spacing rules for the lack of space[3,4].

The current Korean morphological disambiguation systems adopt mainly statistical methods and some of them use rules in the postprocess to filter out incompatible strings[5,6,7]. The statistical approach can, though, give only positive clues of adjacency of two morphemes and it fails to predict precisely their mutual exclusions which play an important role in morphological disambiguation.

In our approach, the morphological analyzer reduces already the number of the candidate morpheme strings using adjacency conditions when it analyses a word into morpheme strings. And then the disambiguation depends on rules and statistics successively. As for the rules, the partial parsing using finite state automata decides the compatibility of each pair of morphemes: a negative value is assigned if a morpheme can not co-occur with another morpheme, while a positive value is given if they are compatible. After applying all the rules related to the word, our system chooses only the positively valued strings. When more than two strings still have same value, the priority in the context is decided by the statistics in the next step.

II. Reduction of Candidate Morpheme Strings

The efficiency and the speed of a morphological disambiguation system depends largely on outputs of a morphological analyzer. When a Korean morphological analyzer segments a word into morpheme strings, one of the serious problems is the over-analysis. To filter out incompatible morpheme strings, our morphological analyzer uses, before the disambiguation, adjacency conditions which can tell the compatibility or the incompatibility of two morphemes. The scope of adjacency conditions in our morphological analyzer is **intra-word**: it concerns the two morphemes which constitute one (compound) word located between two blanks. The adjacency conditions on morpheme pairs can be described either by constraints or by lists: constraints determine their compatibility if any one of the morpheme pairs has strong generative power such as (1-a,b); if its distribution is restricted, we list all its morphemes-pairs as in (2-a,b).

(1) (a) verb stem + ***ki***(nominalization affix) : 먹기, 자기, 공부하기, ...
(b) region name + ***mal***(language[noun]) : 서울말, 미국말, 자카르타말,

(2) (a) *pullyang*(bad[noun]) + ***bae***(group of persons[noun]) : 불량배, 간신배
(b) *twnglok*(registration[noun]) + ***jwng***(certificate[noun]) : 등록증, 영수증

For example, one word *"no-dong-ja-ga"* has a high possibility to produce ambiguities which can be resolved only by semantic analysis.

(3) *no-dong-ja-ga* 노동자가
(a) *no-dong-ja* (laborer[noun]) + *ga* (subject marker[postposition])
(b) **no-dong* (labor[noun]) + *ja-ga* (private house[noun])

If the adjacency condition concerning the noun *"ja-ga* (private house[noun])" tells that its distribution is restricted to the first place of a compound noun, we can output only (3-a) from *"no-dong-ja-ga"* and filter out (3-b). Our morphological analyzer contains adjacency conditions about 1,700 nouns which might have this type of ambiguities.

The word type "one syllable noun + case marker" causes one of the most frequently occurred ambiguities. The example (4) *"su-lwl"* can be analyzed as (4-a) and (4-b).

(4) *su-lwl* 수를
(a) *su* (number[noun]) + *lwl* (object marker[postposition])
(b) **su* (reliance noun) + *lwl* (object marker[postposition])

With the adjacency condition that *"su"* as a reliance noun cannot co-occur with an object marker since it should be followed by an intransitive verb, our system generates only (4-a) from *"su-lwl"* and it filters out (4-b).

The spelling errors that italiqueoccur frequently must also be corrected by the morphological analyzer. According to Korean orthography, a space must be given between *"su*(number[noun])" and its preceding noun. But a large number of examples from our corpus violate this rule. In the example (5), the current Korean morphological analyzers might generate only (5-a) from *"no-dong-ja-su"* which is incorrect.

(5) *#no-dong-ja-su* 노동자수 (#: violation of spacing rule)
(a) **no-dong* (labor[noun]) + *ja-su* (embroidery / self-surrender[noun])
(b) *no-dong-ja* (laborer[noun]) + *su* (number[noun])

Another type of our adjacency conditions gives priority to the parts-of-speech containing noun suffixes such as *"-ja"*, *"-ga"*, etc. followed by *"su"*, even though the word is orthographically incorrect. Our system not only generates (5-b) from *"no-dong-ja-su"*, but also assigns a positive value on it.

Compared to the other systems, the number of candidate morpheme strings can be reduced in our system with those adjacency conditions: when our system analyses successfully more than 99% of the corpus, 33.2% of its outputs show ambiguities and the average ambiguity number per one ambiguous word is 2.75. That means the average number of candidate morpheme strings is only 1.58 per one word.

III. Disambiguation by Rule-Based Approach

In the case that a word still has ambiguities after being processed by the morphological analyzer, our system depends on three different types of the rules for the disambiguation. The scope of those rules is **inter-word**: they can tell the compatibility of a morpheme with its preceding and/or following words. The rules are focused on a governing morpheme and parse its left and right context. Even though the window size of context is determined by the linguistic constraints concerning that morpheme, users can reduce the window size to speed up the disambiguation process.

The first type of the rules concerns 63 specific morphemes (or parts-of-speech) such as *"dae-ha-da* (its conjugated forms *dae-han, dae-hae, dae-hayeu* : be over against)", *"han* (one/ done/ heart-burning)" and *"su* (number/ reliance noun)", since we found that more than 27% of all the ambiguous words in our corpus are related only to those 63 morphemes in some way. After describing the rules to disambiguate those morphemes and/or their related morphemes, we test those rules with a large corpus and refine them. These rules assign a positive or a negative values to the link between each of those morphemes and its following and/or preceding morphemes.

The second type of rules is related to the syntactic constraints of morphemes as follows[8]:

(i) Reliance nouns must follow a word of an adjective form;
(ii) A word of an adjective form should be followed by a noun or another word of an adjective form;
(iii) An intransitive verb or on adjective may not follow a noun of objective form, if they have also a adjective form;
(iv) Declarative endings should precede a quotation mark;
etc.

Each of the rules has different constraint power that affects the positive value of the morpheme when the rule is satisfied. The rule (i) is so strong that the system assigns a high value whereas the rule (iii), less strong than the rule

(i), gives a relatively low value. But a negative value is assigned when any of those rules is not satisfied.

The third type of rules uses the collocation of morphemes. The stem *"ssw"* can have two different categories : a verb (write) and an adjective (bitter). In the case that its subject is *"yak* (medicine[noun])", a plant, etc., it must be an adjective. But if its subject is a human, it must be a verb. We are now trying to make such kind of heuristic rules for the disambiguation, depending mainly on 200,000 different words with high frequency. Those 200,000 words covers 87% of our corpus which contains 11 million words.

The following shows the process of disambiguation explained above in our system.

(A) The result of Morphological Analysis

gol [*gol* (goal[noun])
gol (snore[verb]) + *l* (adjective maker[ending])

su-nwn [*su*^{1} (number[noun]) + *nwn* (topic marker[postposition])
su^2 (reliance noun) + *nwn* (topic marker[postposition])

iss-ta [*iss* (exist[intransitive verb]) + *ta* (declarative maker[ending])

morw-ji [*morw* (do not know[transitive verb]) + *ji* (interrogative marker[ending])

(B) The Process of Disambiguation

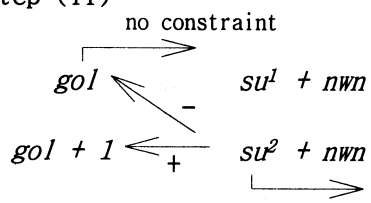
step (I)

gol —(no constraint)→

gol + l —→ The next word must be a noun or a word of an adjective form.

In the first step, the morpheme *"gol"* does not give any constraint on the following morphemes, while the rule (ii) tells that *"gol+l"*, as an adjective form, needs to be followed by a noun or a word of an adjective form.

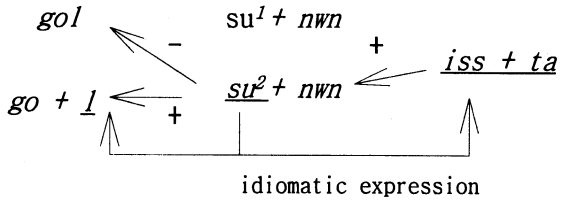
step (II)



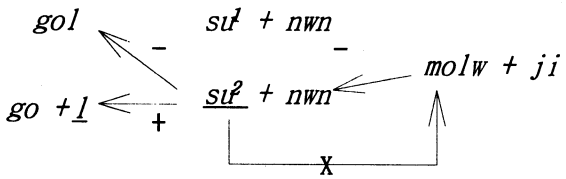
According to the rule (i), "su^2" as a reliance noun requires that the preceding word must be an adjective form. Thus "*gol*" has a negative value with respect to "su^2 + *nwn*" while "*gol* + *l*" has a positive value. In the step (II), the system chooses only "*gol-l* su^2*-nwn*" when any other words are not followed.

Suppose that "*gol-l su-nwn*" is followed by other words. Another constraint concerning "su^2" tells that its following verbs must be "*iss* (exist[intransitive verb])" or "*ebs* (do not exist[intransitive verb])".

step (III-1)



step (III-2)



Since "*-l* su^2+postposition *iss* (can)", defined as an idiomatic expression, assigns a positive value to the link between "su^2*-nwn*" and "*iss-ta*", the highest priority will be given to "*gol-l* su^2*-nwn iss-ta* (can snore)" in the step (III-1). And the same rule gives a negative value to the pair "su^2*-nwn*" and "*mo-lw-ji*". In the step (III-2), "*gol* su^1*-nwn mo-lw-ji* (do not know the number of goals)" is the unique output, because our system filters out all the negatively valued morpheme strings.

IV. Selection of the Most Feasible Morpheme String by Statistics

When we fail to disambiguate a word by applying rules using collocation of morphemes, our system selects the most feasible morpheme string depending on the properties of morphemes constituting one word. Although both the morphological analyzer and the selection routine use only the intra-word information, the selection routine is different from the morphological analyzer in the sense that morphological analyzer uses the constraints in order to remove incorrect morpheme strings whereas the selection depends on the property information for giving priority order to the remaining morpheme strings.

The selection routine uses, as intra-word properties, ① the frequency of a morpheme in the corpus, ② the patterns of a morpheme string according to categories of its morphemes, and ③ special morphemes that can affect the selection of a morpheme string.

In the begining of our research, we expected that the following evaluation function based on the frequency could select effectively the most feasible morpheme string.

$$(6)\quad F(M_1M_2...M_n)=\prod_{i=1}^{n}P(M_i)$$

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

But this function does not work as general heuristics for the following two reasons.

Firstly, the high frequency does not always guarantee 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, since "*i-geul*(this + object marker)", "*jeu-geul*(that + object marker)", "*i-geun*(this + topic marker)", "*jeu-geun*(that + topic marker)", ... occur frequently in the corpus. The high frequency of those morphemes was unexpected. Actually, only a small number of words co-occur with "*-l*(postposition)" or "*-n*(postposition)". Even though their frequency in the corpus is high, most of them are unambiguous.

With the function described in (6), the system selected "*ye-bu* (whether[noun]) + *n*(topic marker)" rather than "*ye-bun* (superfluity [noun])" which is correct. But in ambiguous cases, the morpheme string without "*-n*(topic marker)" and "*-l*(object marker)" has high priority to be corrected.

Secondly, the evaluation value of the morpheme strings containing a

morpheme that does not appear in the our corpus becomes 0. For example, *"so-jil"* can be analyzed as (7-a) and (7-b). In the case that the corpus does not contain(7-b), a morpheme string with *"-l*(object marker)" is selected even if the distribution of (7-a) is extremely restricted. The morpheme string (7-a) is correct only when *"ha-da*(do[verb])" follows it and this case is covered by the collocation rules.

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

In the case of *"gam-gag-gi-do"* which has two interpretations, only (8-a) is selected by the evaluation function, because *"gam-gag-gi*(sensor)" does not exist in the corpus with 11 million words. Although (8-a) is not semantically correct, our morphological analyzer cannot remove this interpretation since it does not use semantic knowledge. The degradation by the missing morphemes in the corpus cannot be ignored to achieve the success rate of the disambiguation more than 97%.

In our experiment, the pattern of a morpheme string using the category information gives better results. Our selection routine prefers "noun + postposition" to "noun + noun" or "noun", if the postposition is neither *"-l*(object marker)" nor *"-n*(topic marker)" and if the noun is not one syllable noun. The morpheme string with one syllable noun is specially dealt with in our system, since they give great difficulties in morphological disambiguation of Korean. For example, *"to-wi*(discussion[noun])" is preferred rather than *"to*(road[noun]) + *wi*(of[postposition])".

The preference patterns are based on the result of the statistical analysis of the patterns in the corpus. In this case, unambiguous words offer the statistical information. The words *"gam-gag-gi-do"* and *"so-jil"* can be successfully disambiguated by the preference patterns. But we use the evaluation function when the patterns are similar. For example, since "verb + ending" and "adjective + ending" are regarded as similar patterns. *"sseu-seu"* is analyzed as both (9-a) and (9-b). As *"ssw*(write[verb])" is used more frequently than *"ssw*(bitter[adjective])", the selection routine gives priority to (9-a) if any other rule disambiguates it.

(9) *"sseu-seu"* 써서
(a) *"ssw*(write[verb]) + *seu*(connecting ending)"
(b) *"ssw*(bitter[adjective) + *seu*(connecting ending)"

The evaluation function also solves the ambiguities caused by the

stemming of a compound noun and a compound predicate. For the disambiguation of the stemming ambiguity, the frequencies of the morphemes in morpheme strings are multiplied. In consequence, the word composed with less morphemes is preferred if all the frequencies of morphemes are same. The preference of less morphemes is a generalized rule which can be adapted to the disambiguation of Korean and Japanese.

If the morphological analyzer fails to analyze a given word or the value of the evaluation value is very low, our system calls the guessing routine for the process of unknown parts of speech. It guesses unknown parts of speech by removing the postposition and the ending attached to the word. If the system guesses more than two different unknown words, it selects the most feasible unknown parts of speech by the frequency of the postpositions and endings.

The accuracy of the guessing routine for the unknown parts of speech is about 98% and that of the disambiguation system is about 97.1% excluding words containing unknown parts of speech.

V. Conclusion

The approach described in this paper is different from the approach which is currently used for the Korean morphological disambiguation in the sense that the rules are applied first and the statistical method is supplementary. Although the rule-based approach is difficult to implement, we may confirm that the accuracy would be improved if we give much more knowledge in the system.

We also assume that making a high quality tagged corpus for Korean is much more difficult than making linguistic rules. By using only linguistic and heuristic rules, we can achieve about 95.3% of accuracy. The accuracy is very high compared to the statistical methods. Until now, any disambiguation system does not exceed 93% of accuracy depending solely on the statistical methods. The processing speed is also not slow, as we adapt demon programming. That is, the dictionary information of the morphemes (or parts of speech) have the rule names to apply the ambiguities related to them. As our system does not use any domain specific rules, it is more robust than the statistical methods, too.

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

References

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

[2] 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*

[3] Seung-Woo Mee. 1994. *New Korean Orthography.* Emunkak:Seoul.

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

[5] 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.*

[6] 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.*

[7] 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.

[8] Nam, Ki-Sim and Young-Geun Go. 1986. *Standard Korean Grammar.* Top Publisher:Seoul.

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

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

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