Structural disambiguation of morpho-syntactic categorial parsing for Korean *

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

Abstract

The Korean Combinatory Categorial Grammar (KCCG) formalism can uniformly handle word order variation among arguments and adjuncts within a clause as well as in complex clauses and across clause boundaries, i.e., long distance scrambling. In this paper, incremental parsing technique of a morpheme graph is developed using the KCCG. We present techniques for choosing the most plausible parse tree using lexical information such as category merge probability, head-head co-occurrence heuristic, and the heuristic based on the coverage of subtrees. The performance results for various models for choosing the most plausible parse tree are compared.

1 Introduction

Korean is a non-configurational, postpositional, agglutinative language. Postpositions, such as noun-endings, verb-endings, and prefinal verb-endings, are morphemes that determine the functional role of NPs (noun phrases) and VPs (verb phrases) in sentences and also transform VPs into NPs or APs (adjective phrases). Since a sequence of prefinal verb-endings, auxiliary verbs and verb-endings can generate hundreds of different usages of the same verb, morpheme-based grammar modeling is considered as a natural consequence for Korean.

There have been various researches to disambiguate the structural ambiguities in parsing. Lexical and contextual information has been shown to be most crucial for many parsing decisions, such as prepositional-phrase attachment (Hindle and Rooth, 1993). (Charniak, 1995; Collins, 1996) use the lexical information and (Magerman and Marcus, 1991; Magerman and Weir, 1992) use the contextual information for structural disambiguation. But, there have been few researches that used probability information for reducing the spurious ambiguities in choosing the most plausible parse tree of CCG formalism, especially for morpho-syntactic parsing of agglutinative language.

In this paper, we describe the probabilistic method (e.g., category merge probability, head-head co-occurrence, coverage heuristics) to reduce the spurious ambiguities and choose the most plausible parse tree for agglutinative languages such as Korean.

2 Overview of KCCG

This section briefly reviews the basic KCCG formalism.

Following (Steedman, 1985), order-preserving type-raising rules are used to convert nouns in grammar into the functors over a verb. The following rules are obligatorily activated during parsing when case-marking morphemes attach to noun stems.

- Type Raising Rules:
  np + case-marker $\Longrightarrow$ $v/(v\backslash$np[case-feature])

This rule indicates that a noun in the presence of a case morpheme becomes a functor looking for a verb on its right; this verb is also a functor looking for the original noun with the appropriate case on its left. After the noun functor combines with the appropriate verb, the result is a functor, which is looking for the remaining arguments of the verb. 'v' is a variable for a verb phrase at any level, e.g., the verb of a matrix clause or the verb of an embedded clause. And 'v' is matched to all of

* This research was partially supported by KOSEF special basic research program (1997.9 $\sim$ 2000.8).

the "v[X]\Args" patterns of the verb categories. Since all case-marked nouns in Korean occur in front of the verb, we don't need to employ the directional rules introduced by (Hoffman, 1995).

We extend the combinatory rules for uncurried functions as follows. The sets indicated by braces in these rules are order-free.

- Forward Application ($\mathsf{A}_>$):
  X/(Args $\cup$ {Y}) Y $\Longrightarrow$ X/Args
- Backward Application ($\mathsf{A}_<$):
  Y X\(Args $\cup$ {Y}) $\Longrightarrow$ X\Args

Using these rules, a verb can apply to its arguments in any order, or as in most cases, the case-marked noun phrases, which are type-raised functors, can apply to the appropriate verbs.

Coordination constructions are modified to allow two type-raised noun phrases that are looking for the same verb to combine together. Since noun phrases, or a noun phrase and adverb phrase, are functors, the following composition rules combine two functions with a set value arguments.

- Forward Composition ($\mathsf{B}_>$):
  X/(X\$Args_X$) Y/(Y\$Args_Y$) $\Longrightarrow$
  X/(X\($Args_X \cup Args_Y$)),
  Y = X\$Args_X$
- Backward Composition ($\mathsf{B}_<$):
  Y\$Args_Y$ X\($Args_X$ $\cup$ {Y}) $\Longrightarrow$
  X\($Args_X \cup Args_Y$)

- Coordination ($\mathbf{\Phi}$):
  X CONJ X $\Longrightarrow$ X

3 Basic morph-syntactic chart parsing

Korean chart parser has been developed based on our KCCG modeling with a 100,000 morpheme dictionary. Each morpheme entry in the dictionary has morphological category, morphotactics connectivity and KCCG syntax categories for the morpheme.

In the morphological analysis stage, a unknown word treatment method based on a morpheme pattern dictionary and syllable bigrams is used after (Cha et al., 1998). POS(part-of-speech) tagger which is tightly coupled with the morphological analyzer removes the irrelevant morpheme candidates from the morpheme graph. The morpheme graph is a compact representation method of Korean morphological structure. KCCG parser analyzes the morpheme graph at once through the morpheme graph embedding technique (Lee et al., 1996).

The KCCG parser incrementally analyzes the sentence, eojeol by eojeol [1]. Whenever an eojeol is newly processed by the morphological analyzer, the morphemes resulted in a new morpheme graph are embedded in a chart and analyzed and combined with the previous parsing results.

4 Statistical structured disambiguation for KCCG parsing

The statistics which have been used in the experiments have been collected from the KCCG parsed corpora. The data required for training have been collected by parsing the standard Korean sentence types[2], example sentences of grammar book, and colloquial sentences in trade interview domain[3] and hotel reservation domain[4]. We use about 1500 sentences for training and 591 independent sentences for evaluation.

The evaluation is based on parseval method (Black et al., 1991). In the evaluation, "No-crossing" is the number of sentences which have no crossing brackets between the result and the corresponding correct trees of the sentences. "Ave. crossing" is the average number of crossings per sentence.

4.1 Basic statistical model

A basic method of choosing the most plausible parse tree is to order the probabilities by the lexical preferences[5] and the syntactic merge probability. In general, a statistical parsing model defines the conditional probability, $P(\tau|S)$, for each candidate tree τ for a sentence S. A generative model uses the observation that maximising $P(\tau, S)$ is equivalent to maximising $P(\tau|S)$[6].

[1] Eojeol is a spacing unit in Korean and is similar to an English word.

[2] Sentences of length ≤ 11.

[3] Sentences of length ≤ 25.

[4] Sentences of length ≤ 13.

[5] The frequency with which a certain category is associated with a morpheme tagged for part-of-speech.

[6] $P(S)$ is constant.

Thus, when S is a sentence consisted of a sequence of morphemes tagged for part-of-speech, $(w_1, t_1), (w_2, t_2), ..., (w_n, t_n)$, where w_i is a i^{th} morpheme, t_i is the part-of-speech tag of the morpheme w_i, and c_{ij} is a category with relative position i, j, the basic statistical model will be given by:

$$\tau^* = arg\max_{\tau} P(\tau|S) \quad (1)$$

$$= arg\max_{\tau} \frac{P(\tau, S)}{P(S)} \quad (2)$$

$$\approx arg\max_{\tau} P(\tau, S). \quad (3)$$

The τ^* is the probabilities of the optimal parse tree.

$P(\tau, S)$ is then estimated by attaching probabilities to a bottom-up composition of the tree.

$$P(\tau, S) = \prod_{c_{ij} \in \tau} P(c_{ij}) \quad (4)$$

$$= \prod_{c_{ij} \in \tau} (P(c_{ij}|c_{ik}, c_{k+1j}) \times P(c_{ik})P(c_{k+1j})), \quad (5)$$

$i \leq k \leq j$,
if c_{ij} *is a terminal*,
then $P(c_{ij}) = P(c_{ij}|w_i, t_i)$,

and

$$P(c_{ij}|t_i, w_i) \approx \frac{frequency(c_{ij}, t_i, w_i)}{frequency(t_i, w_i)}, \quad (6)$$

$$P(c_{ij}|c_{ik}, c_{k+1j}) \approx \frac{frequency(c_{ij}, c_{ik}, c_{k+1j})}{frequency(c_{ik}, c_{k+1j})}. \quad (7)$$

The basic statistical model has been applied to morpheme/part-of-speech/category 3-tuple. Due to the sparseness of the data, we have used part-of-speech/category pairs[7] together, i.e., collected the frequencies of the categories associated with the part-of-speeches assigned to the morpheme. Table 1 illustrates the sample entries of the category probability database. In table, 'nal (fly)' has two categories with 0.6375 and 0.3625 probability respectively. Table 2 illustrates the sample entries of the merge probability database using equation 7.

[7]We define this as $P(c_{ij}|t_i) \approx \frac{frequency(c_{ij}, t_i)}{frequency(t_i)}$.

Table 3: Results from the Basic Statistical Model

Total sentences	591
No-crossing	74.62%
Ave. crossing	1.00
Labeled Recall	77.02
Labeled Precision	79.15

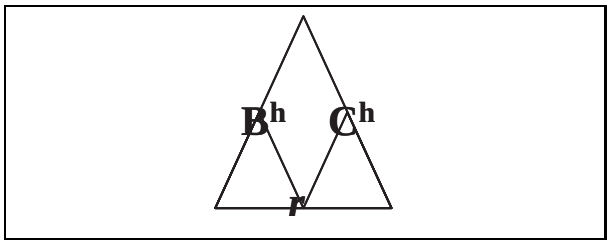

Figure 1: Sub-constituents for head-head co-occurrence heuristics

Table 3 summarizes the results on an open test set of 591 sentences.

4.2 Head-head co-occurrence heuristics

In the basic statistical model, lexical dependencies between morphemes that take part in merging process cannot be incorporated into the model. When there is a different morpheme with the same syntactic category, it can be a miss match on merging process. This limitation can be overcome through the co-occurrence between the head morphemes of left and right sub-constituent.

When B^h is a head morpheme of left sub-constituent, r is a case relation, C^h is a head morpheme of right sub-constituent as shown in figure 1, head-head co-occurrence heuristics are defined by:

$$P(B^h|r, C^h) \approx \frac{frequency(B^h, r, C^h)}{frequency(r, C^h)}. \quad (8)$$

The head-head co-occurrence heuristics have been augmented to equation 5 to model the lexical co-occurrence preference in category merging process. Table 4 illustrates the sample entries of the co-occurrence probability database. In Table 4, a morpheme 'sae (means 'bird')', which has a "MCK (common noun)" as POS tag, has been used a nominative of verb 'nal (means 'fly')' with 0.8925 probability.

Table 1: Sample entries of the category probability database ('DIl' means an 'l' irregular verb.)

POS, morpheme	category	probability
DIl, nal	v[D]\{np[nom]}	0.6375
DIl, nal	v[D]\{np[nom],np[acc]}	0.3625
DIl	v[D]\{np[nom]}	0.3079
DIl	v[D]\{np[nom],np[acc]}	0.2020

Table 2: Sample entries of syntactic merge probability database

left category	right category	merged category	probability
$v/(v\backslash$np[nom])	v[D]\{np[nom],np[acc]}	v[D]\{np[acc]}	0.0473
$v/(v\backslash$np[acc])	v[D]\{np[nom],np[acc]}	v[D]\{np[nom]}	0.6250
np	$(v/(v\backslash$nom))\np	$v/(v\backslash$np[nom])	0.2197

The modified model has been tested on the same set of the open sentences as in the basic model experiment. Table 5 summarizes the result of these experiments.

- *Experiment: (linear combination of the basic model and the head-head co-occurrence heuristics).*

$$
\begin{aligned}
P(\tau, S) \quad = \quad & \prod_{c_{ij} \in \tau} ((\alpha P(c_{ij}|c_{ik}, c_{k+1j}) \\
& + \beta P(B^h|r, C^h)) \\
& \times P(c_{ik}) P(c_{k+1j})), \qquad (9) \\
& i \leq k \leq j, \\
& \textit{if } c_{ij} \textit{ is a terminal}, \\
& \textit{then } P(c_{ij}) = P(c_{ij}|w_i, t_i).
\end{aligned}
$$

Table 5: Results from the Basic Statistical Model plus head-head co-occurrence heuristics

Total sentences	591
No-crossing	81.05%
Ave. crossing	0.70
Labeled Recall	84.02
Labeled Precision	85.30

4.3 The coverage heuristics

If there is a case relation or a modification relation in two constituents, coverage heuristics designate it is easier to add the smaller tree to the larger one than to merge the two medium sized trees. On the contrary, in the coordination relation, it is easier to merge two medium sized trees. We implemented these heuristics using the following coverage score:

Case relation, modification relation:

$$
COV\_score = \frac{left\ subtree\ coverage + right\ subtree\ coverage}{4 \times \sqrt{left\ subtree\ coverage \times right\ subtree\ coverage}} \qquad (10)
$$

Coordination:

$$
COV\_score = \frac{2 \times \sqrt{left\ subtree\ coverage \times right\ subtree\ coverage}}{left\ subtree\ coverage + right\ subtree\ coverage} \qquad (11)
$$

A coverage heuristics are added to the basic model to model the structural preferences. Table 6 shows the results of the experiments on the same set of the open sentences.

- *Experiment: (the basic model to the COV_score heuristics).* We have used the COV_score as the exponent weight feature for this experiment since the two numbers are in the different nature of statistics.

$$
\begin{aligned}
P(\tau, S) \quad = \quad & \prod_{c_{ij} \in \tau} (P(c_{ij}|c_{ik}, c_{k+1j})^{1-COV\_score} \\
& \times P(c_{ik}) P(c_{k+1j})), \qquad (12) \\
& i \leq k \leq j, \\
& \textit{if } c_{ij} \textit{ is a terminal}, \\
& \textit{then } P(c_{ij}) = P(c_{ij}|w_i, t_i).
\end{aligned}
$$

Table 4: Sample entries of co-occurrence probability database.

head-head co-occurrence	probability
(MCC<ganeungseong>,np[nom],HR<nob>)	0.8932
(MCK<sae>,np[nom],DIl<nal>)	0.8925
(MCK<galeuchim>,np[acc],DIeu<ddaleu>)	0.8743

Table 6: Results from the Basic Statistical model plus Coverage heuristics

Total sentences	591
No-crossing	80.13%
Ave. crossing	0.81
Labeled Recall	82.59
Labeled Precision	83.75

5 Summary

We developed a morpho-syntactic categorial parser of Korean and devised a morpheme-based statistical structural disambiguation schemes.

Through the KCCG model, we successfully handled difficult Korean modeling problems, including relative free-word ordering, coordination, and case-marking, during the parsing.

To extract the most plausible parse trees from the parse forest, we have presented basic statistical techniques using the lexical and contextual information such as morpheme-category probability and category merge probability.

Two different nature of heuristics, head-head co-occurrence and coverage scores, are also developed and tested to augment the basic statistical model. Each of them demonstrates reasonable performance increase.

The next step will be to devise more heuristics and good combination strategies for the different nature of heuristics.

References

E. Black, S. Abney, D. Flickenger, C. Gdaniec, R. Grishman, P. Harrison, D. Hindle, R. Ingria, F. Jelinek, J. Klavans, M. Liberman, M. Marcus, S. Roukos, B. Santorini, and T. Strzalkowski. 1991. A Procedure for Quantitatively Comparing the Syntactic Coverage of English Grammars. In *Proc. of Fourth DARPA Speech and Natural Language Workshop*.

Jeongwon Cha, Geunbae Lee, and Jong-Hyeok Lee. 1998. Generalized unknown morpheme guessing for hybrid pos tagging of korean. In *Proceedings of Sixth Workshop on Very Large Corpora in Coling-ACL 98*, Montreal, Canada.

E. Charniak. 1995. Prsing with Context-Free Grammars and Word Statistics. Technical Report CS-95-28, Brown University.

M. Collins. 1996. A New Statistical Parser Based on Bigram Lexical Dependencies. In *Proceedings of the 34th Annual Meeting of the ACL, Santa Cruz.*

D. Hindle and M. Rooth. 1993. Structural ambiguity and lexical relations. *Computational Linguistics*, 19(1):103–120.

B. Hoffman. 1995. *The Computational Analysis of the Syntax and Interpretation of "Free" Word Order in Turkish.* Ph.D. thesis, University of Pennsylvania. IRCS Report 95-17.

Wonil Lee, Geunbae Lee, and Jong-Hyeok Lee. 1996. Chart-driven connectionist categorial parsing of spoken korean. *Computer processing of oriental languages*, Vol 10, No 2:147–159.

D. M. Magerman and M. P. Marcus. 1991. Parsing the voyager domain using pearl. In *In Proc. Of the DARPA Speech and Natural Language Workshop*, pages 231–236.

D. M. Magerman and C. Weir. 1992. Efficiency, robustness and accuracy in picky chart parsing. In *In Proc. Of the 30th Annual Meeting of the Assoc. For Computational Linguistics(ACL-92)*, pages 40–47.

Mark Steedman. 1985. Dependency and Coordination in the Grammar of Dutch and English. *Language*, 61:523–568.