Decision Tree-Based Noun Phrase Detection and Classification in Agglutinative Languages

Photina Jaeyoun Jang and Christopher Hogan and Robert E. Frederking

January 25, 1999

Paper ID Code: 136
Submission Type: Thematic Session
Topic Areas or Theme ID: L1
Word Count: 2589
Under consideration for other conferences (specify)? None.

Abstract

The current paradigm in parsing has been developed primarily using English, a language that relies on word order to express grammatical function. However, most languages in the world rely much more on NP-marking to express the same functions. We propose therefore a shallow NP parsing technique which makes much more use of NP-marking, and evaluate the technique on Korean, an agglutinating language.

1 Introduction

In this paper, we take a shallow parsing approach to identifying Noun Phrases and their grammatical relationship to the verb. Unlike more conventional parsers, our techniques rely on local analysis and the use of decision trees which are trained on syntactically annotated corpora.

1.1 Coding Grammatical Function

One of the most important reasons for performing parsing is to determine the functional relationships that exist between constituents in a sentence. Determining these relationships is necessary for many, if not most, applications of parsing, including Machine Translation and

Natural Language Understanding. Historically, among the most useful of these relationships has been that between the verb and the noun phrases which serve as its arguments, also known as *semantic, thematic* or *grammatical roles.*[1]

There are three techniques which languages use to express grammatical function (Andrews, 1992)

1. *Order and arrangement*

   Grammatical function is indicated by the relative order of the NPs relative to the verb.

2. *NP-marking*

   Grammatical function is indicated by a morphological marker on the NP.

3. *Cross-referencing*

   Grammatical function is indicated by marking some grammatical property of the NP on the verb.

Although English for the most part uses the order and arrangement strategy to indicate grammatical function, this is by no means common among the world's languages. Indeed, no language makes exclusive use of ordering, while every language uses NP-marking to some extent (Andrews, 1992, p. 72). One consequence of this is that languages which rely primarily on a non-order-based coding strategy allow freer word order than those which use order as an indicator of grammatical function. Word order in such languages often expresses pragmatic, rather than grammatical function.

The following examples illustrate the freer word order exhibited by languages that code gram-

[1] Typically, the term *grammatical role* refers to surface relationships, such as SUBJECT, OBJECT, INDIRECT OBJECT, *etc.*, while *semantic role* or *thematic role* refer to the underlying semantic relationships, such as AGENT, THEME, GOAL, *etc.* The mapping between these two groups is very often not one-to-one.

matical function using NP-marking. Korean is an agglutinating language that permits free reordering of NP and Verb constituents in main clauses, so-called NP-scrambling (Choe, 1989).[2] The following sentences are semantically equivalent, differing only in pragmatic effect, *e.g.*, which constituent receives focus (Lee, 1989).

(1) *gae.ga* *dalg.eul* *bo.n.da*
dog.SUBJ hen.OBJ see.PRES.DECL
'A dog looks at a hen'

(2) *gaega* *bonda* *dalgeul*
SUBJ VERB OBJ

(3) *dalgeul* *bonda* *gaega*
OBJ VERB SUBJ

(4) *bonda* *dalgeul* *gaega*
VERB OBJ SUBJ

Besides Korean, scrambling is exhibited by many languages which use NP-marking as their primary coding strategy, including Japanese, German and Warlpiri, among others (Marácz and Muysken, 1989). These are languages where conventional parsing techniques, which rely to a good extent on word order, will encounter difficulties. We therefore believe that it is necessary to investigate parsing techniques that make greater use of NP-marking and cross-referencing, and less use of order and arrangement.

1.2 Shallow Parsing

Shallow parsing is a generic term for techniques which produce less comprehensive parses than those obtained from conventional parsers (Karlsson and Karttunen, 1996). Such techniques are

[2] Note that the freedom of word order only applies between NP and verb constituents. The word order within these constituents remains fixed, although it is not obvious from these examples.

desirable in cases where conventional parsers fail, or where simpler analyses are adequate. In addition, shallow parsers, where usable, may be much more easily and rapidly built than full parsers.

It is possible to compare techniques for shallow parsing with respect to their sensitivity to word order. While shallow parsers often succeed where full parsers fail, most shallow parsers have been developed (as have most full parsers) for English and other languages in which word order is fairly fixed. Thus, techniques based on Context Free Grammars (Hindle, 1989; Joshi and Srinivas, 1994) or (Stochastic) Finite State Grammars (Manning, 1993; Ejerhed and Church, 1983; Koskenniemi, 1990; Aït-Mokhtar and Chanod, 1997; Senellart, 1998) may suffer from the same problems as do full parsers, as may any technique that is based primarily on recognizing sequences of words or tags (Hobbs et al., 1993; Schwarz, 1990). In contrast, techniques which make relatively little use of sequences of words or tags should perform more capably when they encounter freer word order. These techniques include local (stochastic) analysis (Church, 1988; Bahl et al., 1991), methods based on function/content word distinctions (Abney, 1994; Ross and Tukey, 1975; Bourigault, 1992; Smith and Witten, 1993), and those based on lexical cohesion (Stolz, 1965; Magerman and Marcus, 1990; Harris, 1955).

Another dimension along which approaches to parsing vary is the degree to which they are automatically learned from data vs. coded by hand. Traditionally, parsers have been developed by producing hand-coded rules, usually based on a native speaker's knowledge of the language (Mitamura et al., 1991) or deduced from data by a human expert (Oostdijk, 1991). While it is possible to produce very high accuracy in this way, such efforts are extremely knowledge intensive. More recently, many approaches have been developed for automatically acquiring the necessary information for parsing from corpora, whether syntactically annotated or not

(Carroll et al., 1996; Brill et al., 1990). Although conventional grammars can be learned (Black et al., 1993), the techniques are more often used for developing shallow parsers, as most of the available corpora are not annotated in detail (Marcus et al., 1993).

In this paper, we present a method for automatically learning to produce NP bracketings for simple Korean sentences, and to tag the bracketings as either subject or object. We describe a method of shallow parsing that is based on decision trees, and present the results of some experiments to evaluate the technique on in- and out-of-domain data.

2 Learning Noun Phrases with Decision Trees

Our approach to shallow parsing relies on machine learning of local context to determine the boundaries of noun phrases, and is thus similar to work by Church (1988). However, whereas Church used a probabilistic method predicated on the sequence of words and part-of-speech tags in the sentence, we have chosen to use decision trees, and to enrich the set of objects to which the learning technique can refer, specifically including morphology, which is vital to properly parsing languages which use NP-marking. In addition, we require that the parser return information about the grammatical function of the identified NPs.

A Decision Tree is a machine learning technique that learns to associate a discrete outcome to a set of pre-identified attributes (Quinlan, 1993). Induction algorithms, such as ID3 and C4.5, exist to take labelled data and produce a decision tree as output (Quinlan, 1986). Decision trees have proven to be very useful for machine learning, and have been applied to Natural Language Processing tasks such as part-of-speech tagging (Stein and Schmid, 1995; Màrquez and guez, 1996), parsing (Magerman, 1995), grapheme-to-phoneme conversion (Daelemans and den Bosch, 1994; Pagel et al., 1998) and morpheme boundary detection (van den Bosch et al., 1995).

2.1 Locating and Labelling Noun Phrases

The basic problem we address here is how to determine the location of noun phrase boundaries in a utterance, and to properly label them as being SUBJECT or OBJECT.[3] To this end, if we are given a sequence of words:

(5) $w_1 \quad w_2 \quad \ldots \quad w_7$

which represent an utterance, we wish to use a trained decision tree to determine which subsequences of words represent noun phrases. If we present each subsequence which could conceivably be a noun phrase to the decision tree,[4] it will answer one of: `subj`, `obj` or `neither`.

(6) $[\ w_1\ ] \Rightarrow$ *neither*

(7) $[\ w_1 \quad w_2\ ] \Rightarrow$ *obj*

(8) $[\ w_1 \quad w_2 \quad w_3\ ] \Rightarrow$ *subj*

$\vdots$

(9) $[\ w_4 \quad w_5 \quad w_6\ ] \Rightarrow$ *obj*

$\vdots$

(10) $[\ w_7\ ] \Rightarrow$ *neither*

These decisions can be used to determine a proper bracketing of the utterance:

[3]In these experiments, we were not concerned with NPs which were neither subjects nor objects.

[4]Although the number of subsequences is in principle quite large, restrictions on the length and constraints such as no crossing branches will reduce this number.

(11) $[\ \ [\ \ w_1\ \ w_2\ \ ]_{obj}\ \ w_3\ \ ]_{subj}\ \ [\ \ w_4\ \ w_5\ \ w_6\ \ ]_{obj}\ \ w_7$

Training proceeds in much the same way, except that a syntactically annotated corpus is used to derive properly labelled training examples, from which certain attributes are extracted, as discussed below. These training examples are then used to learn a decision tree.

2.2 Attributes

We identified twelve attributes for the decision trees. Because we were uncertain which attributes might prove useful, we supplied all twelve to the decision tree induction algorithm, which then chose the most informative ones. In order to increase accuracy, a preprocessing step consisting of morphological segmentation and part-of-speech tagging was first performed on the utterances, and certain attributes refer to this information.

1. *tag sequence*

   The sequence of tags from the part-of-speech tagger that make up the NP.

2. *case marker*

   The case marker of the head noun as determined by morphological analysis.

3. *count of modifiers*

   The number of modifiers of the head noun (*e.g.*, adjectives, adverbs, *etc.*) present in the NP.

4. *count of noun heads*

   Number of noun heads present in the NP.

5. *last syllable*

   Identity of the final syllable of the NP.

6. *count of words*

   Number of words in the NP.

7–12. *words*

   Identity of the 1st–6th words in the NP, or `NONE` if the NP is less than 6 words long.

3 Experiments

In this section, we describe several experiments we performed with the setup described earlier.

3.1 Data

As mentioned earlier, it is necessary to train the decision trees on syntactically annotated corpora. Unfortunately, although many corpora exist for English, and some are in development for other European languages, there are no commonly available corpora for Korean.[5] Furthermore, our application was conversational speech, which is rather distinct from the language of written materials, and for which there exist very few corpora, much less syntactically annotated ones.

In order to circumvent these problems, we selected two near-in-domain texts, and annotated them in-house.[6] The sizes of these corpora are shown below.

	UTTERANCES	WORD TOKENS
middle school	426	1598
recording	730	1367

[5] It appears now that at least one group in Korea is currently producing such corpora (Chang et al., 1997).

[6] We are indebted to Soyoung Kang for her effort in labelling our corpora.

The first of these corpora, the middle school corpus, includes conversations between participants[7] at a middle school in Korea. The speech is simple, but formal, and the transcription has been cleaned to ensure that the utterances are grammatical. The second corpus, the recording corpus, was produced by recording a group of Korean college students in conversation. The speech is both simple and colloquial, and all disfluencies and ungrammatical utterances have been accurately transcribed.

The corpora were annotated with a simple scheme that indicated phrasal boundaries and identified their grammatical functions. Typical examples of the corpora and annotation scheme are given below.

(12) $[\ \textit{nanun}\ ]_{subj}\ [\ \textit{yakwulul}\ ]_{obj}\ [\ \textit{choahae}\ ]_{pred}$

'I like baseball'

(13) $[\ \textit{nanun}\ ]_{subj}\ [\ \textit{paykopha}\ ]_{pred}\ [\ \textit{mwuel}\ ]_{subj}\ [\ \textit{mekci}\ ]_{pred}$

'I am hungry. What can I eat?'

3.2 Preprocessing

As mentioned earlier, preprocessing was performed to increase the accuracy of the decision trees. As a preprocessor, we used a publicly available application (Lee et al., 1995) that performed morphological analysis and part-of-speech tagging. Additional preprocessing was performed to extract noun phrases, and to produce the attributes and predicted variables for decision tree training.

[7]Including teachers, students, and family members.

3.3 Training

For decision tree induction, we used the C4.5 algorithm (Quinlan, 1993). The algorithm includes a pruning step, which holds out a small amount of the training data. This data is used to remove certain branches of the decision tree. The pruning step is intended to reduce the problem of overfitting, or too accurately modeling the training data at the expense of novel data. Pruning typically improves the accuracy of the decision tree by a few percentage points.

Following standard procedure, the preprocessed data were divided randomly into a training set and a testing set. We performed tests with both a 75%-25% and a 66%-33% training-test set split.

3.4 Results

The first aspect of the system we wished to test was whether or not it could correctly classify pre-bracketed noun phrases. While this is not a significantly difficult test, it establishes that the technique can work. For this set of tests, we used only the middle school corpus, and all twelve attributes described above. The preprocessed corpus consisting of 1700 NPs was run with both 66%-33% and 75%-25% split. Each test was performed several times with different random splits of the corpus. Error rates on the test sets are given below. A baseline, based on the strategy of "always guess `NONE`"[8] is also given for comparison.

	run 1	run 2	run 3	run 4	average error	baseline error
66%-33% split	3.5%	2.8%	3.5%	—	3.3%	21.5%
75%-25% split	2.6%	3.8%	3.5%	3.5%	3.4%	22.6%

[8] `NONE` is the most prevalent label in the data.

Since the decision trees are clearly able to correctly label NPs accurately, the next test concerned their ability to properly detect NP boundaries. This test was again carried out using only the middle school corpus. Examination of the decision trees produced in the previous step suggested that not all of the attributes were equally useful. In fact, 7 out of the 12 attributes were contributing most highly to overtraining and were pruned away by the algorithm. Thus, we have only used attributes 2–6 on this and the remaining tests. Again, error rates are reported.

	run 1	run 2	run 3	run 4	average
75%-25% split	5.26%	8.03%	6.73%	9.75%	7.44%

Finally, we tested the decision trees trained on the middle school corpus on the recording corpus, in order to determine the degree to which accuracy depends on domain-specific features (recall that the recording corpus is much more colloquial than the middle school corpus). We measured recall (number of NPs found / total NPs in test set) and precision (number of correctly classified NPs / number of NPs found).

	run 1
recall	62.58%
precision	91.41%

3.5 Discussion

The results show that the decision-tree based noun phrase bracketer performs quite well, at least in-domain. The first test demonstrates that the algorithm is able to do very well at labelling NPs that are presented to it. The second test indicates that it is also able to do quite well at bracketing NPs in sentences. In both of these tests, the decision tree was able to draw on morphological information present in the data to properly determine the location and type of the NPs.

The attempt to transfer this data into a different domain did not do as well. While precision remained high, recall was fairly low. We believe the main reason for this was the different properties of the two corpora. Although both represented spoken language, the middle school text had been cleaned up so as to be more grammatical, while the recording corpus had not been. Among the differences was that the morphological case markers, which are were among the most informative attributes in the middle school corpus, were often missing in the recording corpus. This is a property of rapid Korean speech. Without this vital information, the decision trees fell back on the 'last syllable' attribute, which proved not to be as effective at predicting phrase boundaries. In addition, the recording corpus had frequent instances of contractions within single words and between words, which resulted in poor performance from the part-of-speech tagger, and interfered with the ability of the decision tree to properly utilize certain attributes.

4 Application

The noun phrase bracketing technique described in this paper has been implemented as a part of the DIPLOMAT rapid-deployment speech-to-speech machine translation system (Frederking et al., 1998), although system testing with the new technique has not yet been carried out. Developing machine translation systems has traditionally been a very time consuming, knowledge-intensive task, often taking years to produce. The DIPLOMAT system explores new technologies which can accelerate this process. The current work is intended to be used in place of a conventional parser which might have difficulties with conversational speech in freer word order languages with one that is more robust and can be developed more quickly.

5 Conclusion

We have presented a technique for identifying and labelling noun phrases that is based on decision trees. Our technique is designed to overcome the difficulties that conventional parsers encounter when dealing with NP-marking languages that exhibit freer word orders. Tests using Korean demonstrate that our technique is effective. Unfortunately, our technique also exhibits a sensitivity to domain specificity, with performance declining significantly when shifted to a more colloquial domain.

References

Steven Abney. 1994. Partial parsing. A tutorial presented at the Fourth Conference on Applied Natural Language Processing (ANLP '94), October.

Salah Aït-Mokhtar and Jean-Pierre Chanod. 1997. Incremental finite-state parsing. In *Proceedings of the Fifth Conference on Applied Natural Language Processing (ANLP '97)*, pages 72–79, Washington, D.C., April.

Avery Andrews. 1992. The major functions of the noun phrase. In Timothy Shopen, editor, *Language typology and syntactic description*, volume I, chapter 2, pages 62–154. Cambridge University Press, Cambridge, England.

Lalit R. Bahl, Peter F. Brown, Peter V. de Souza, and Robert L. Mercer. 1991. A tree-based statistical language model for natural language speech recognition. *IEEE Transactions on Acoustics, Speech and Signal Processing*, pages 507–514.

E. Black, R. Garside, and G. Leech, editors. 1993. *Statistically-Driven Computer Grammars of English: The IBM/Lancaster Approach*. Rodopi, Amsterdam.

Didier Bourigault. 1992. Surface grammatical analysis for the extraction of terminological noun phrases. In *Proceedings of the 14th International Conference on Computational Linguistics (COLING '92)*, volume III, pages 977–981.

Eric Brill, David Magerman, Mitchell P. Marcus, and Beatrice Santorini. 1990. Deducing linguistic structure from the statistics of large corpora. In *Proceedings of the DARPA Speech and Natural Language Workshop*, pages 275–282.

John Carroll, Ted Briscoe, Nicoletta Calzolari, Stefano Federici, Simonetta Montemagni, Vito Pirrelli, Greg Grefenstette, Antonio Sanfilippo, Glenn Carroll, and Mats Rooth. 1996. Sparkle work package 1: Specification of phrasal parsing prefinal report. Available online at: http://www.ilc.pi.cnr.it/sparkle/wp1-prefinal/wp1-prefinal.html, May.

Byung Gyu Chang, Kong Joo Lee, and Gil Chang Kim. 1997. Design and implementation of tree tagging workbench to build a large tree tagged corpus of Korean. In *Proceedings of the 9th Conference on Hangul and Korean Language Information Processing*, pages 421–429, Pusan, Korea.

Hyon Sook Choe. 1989. Restructuring parameters and scrambling in Korean and Hungarian. In László Marácz and Pieter Muysken, editors, *Configurationality: The Typology of Asymmetries*, number 34 in Studies in Generative Grammar, pages 267–292. Foris Publications, Dordrect, Holland.

Kenneth Church. 1988. A stochastic parts program and noun phrase parser for unrestricted text. In *Proceedings of the Second Conference on Applied Natural Language Processing*, pages 136–143, Austin, Texas. ACL.

W. Daelemans and A. Van den Bosch. 1994. A language-independent, data-oriented architecture for grapheme-to-phoneme conversion. In *Proceedings of ESCA-IEEE Speech Synthesis Conference '94*, New York.

Eva Ejerhed and Kenneth Church. 1983. Finite state parsing. In Fred Karlsson, editor, *Papers from the Seventh Scandinavian Conference of Linguistics*, pages 410–432.

Robert Frederking, Alexander Rudnicky, and Christopher Hogan. 1998. Interactive speech translation in the DIPLOMAT project. *Machine Translation.* (submitted).

Zellig Harris. 1955. From phoneme to morpheme. *Language*, 31.

Donald Hindle. 1989. Acquiring disambiguation rules from text. In *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics (ACL '89)*, pages 118–125, Vancouver, Canada.

Jerry R. Hobbs, Douglas Appelt, Mabry Tyson, and Megumi Kameyama. 1993. FASTUS: A system for extracting information from text. In *ARPA Workshop on Human Language Technology*, San Mateo, California. Defense Advanced Research Projects Agency, Morgan Kaufmann.

Aravind K. Joshi and B. Srinivas. 1994. Disambiguation of super parts of speech (or supertags): Almost parsing. In *Proceedings of the 15th International Conference on Computational Linguistics (COLING '94)*, Kyoto, Japan, August.

Fred Karlsson and Lauri Karttunen. 1996. Sub-sentential processing. In Ronald A. Cole et al., editors, *Survey of the State of the Art in Human Language Technology*, chapter 3.2. Center for Spoken Language Understanding, Oregon Graduate Institute, USA. Available online at http://cslu.cse.ogi.edu/HLTsurvey/ch3node4.html.

Kimmo Koskenniemi. 1990. Finite-state parsing and disambiguation. In H. Karlgren, editor, *Proceedings of the 13th International Conference on Computational Linguistics (COLING '90)*, pages 229–232.

Sangho Lee, Junyun Seo, and Yung-Hwan Oh. 1995. A Korean part-of-speech tagging system with handling unknown words. In *Proceedings of the 1995 International Conference on Computer Processing of Oriental Languages*, pages 164–171, Honolulu, Hawaii, November. Available online at http://bulsai.kaist.ac.kr/~shlee/ktsworld.htm.

Hansol H. B. Lee. 1989. *Korean Grammar.* Oxford University Press, Oxford, England.

D. Magerman and M. Marcus. 1990. Parsing a natural language using mutual information statistics. In *Proceedings of the Eighth National Conference on Artificial Intelligence (AAAI '90)*, August.

David M. Magerman. 1995. Statistical decision-tree models for parsing. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics (ACL '95)*, Cambridge, Massachussets, June.

Christopher D. Manning. 1993. Automatic acquisition of a large subcategorization dictionary from corpora. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics (ACL '93)*, pages 235–242.

László Marácz and Pieter Muysken, editors. 1989. *Configurationality: The Typology of Asymmetries.* Number 34 in Studies in Generative Grammar. Foris Publications, Dordrecht, Holland.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn treebank. *Computational Linguistics*, 19(2):313–330.

Lluis Màrquez and Horacio Rodríguez. 1996. Towards learning a constraint grammar from annotated corpora using decision trees. Technical Report LSI-96-9-R, Universitat Politècnica de Catalunya: Departament de Llenguatges i Sistemes Informàtics.

T. Mitamura, E. Nyberg, and J. Carbonell. 1991. An efficient interlingua translation system for multi-lingual document production. In *Proceedings of the Third Machine Translation Summit*, Washington, D.C.

N. Oostdijk. 1991. *Corpus Linguistics and the Automatic Analysis of English.* Rodopi, Amsterdam.

Vincent Pagel, Kevin Lenzo, and Alan Black. 1998. Letter to sound rules for lexicon compression. In *Proceedings of the Fifth International Conference on Spoken Language Processing (ICSLP 99)*, Sydney, Australia.

J. Quinlan. 1986. Induction of decision trees. *Machine Learning*, 1:81–206.

J. Ross Quinlan. 1993. *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers, Inc., San Mateo, California.

Ian C. Ross and John W. Tukey. 1975. Introduction to these volumes. In *Index to Statistics and Probability*, pages iv–x. The R & D Press, Los Altos, California.

Christoph Schwarz. 1990. Automatic syntactic analysis of free text. *Journal of the American Society for Information Science*, 41(6):408–417.

Jean Senellart. 1998. Locating noun phrases with finite state transducers. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and the 17th International Conference on Computational Linguistics (COLING-ACL '98)*, volume II, pages 1212–1219, Montreal, Quebec, Canada, August.

Tony C. Smith and Ian H. Witten. 1993. Language inference from function words. ms., University of Calgary and University of Waikato, January.

Achim Stein and Helmut Schmid. 1995. Etiquetage morphologique de textes français avec un arbre de décisions. *Traitement automatique du langage*, 36(1–2).

W. Stolz. 1965. A probabilistic procedure for grouping words into phrases. *Language and Speech*, 8:219–235.

A. van den Bosch, W. Daelemans, and A. Weijters. 1995. Machine learning of morpheme boundary detection. In *Abstract for the Sixth Computational Linguistics in the Netherlands Meeting (CLIN '95)*, page 35, Antwerpen.