Hypothesis Selection in Grammar Acquisition

Masaki KIYONO* and **Jun'ichi TSUJII**
Centre for Computational Linguistics
University of Manchester Institute of Science and Technology
PO Box 88, Manchester M60 1QD
United Kingdom
kiyono@ccl.umist.ac.uk, tsujii@ccl.umist.ac.uk

Abstract

This paper presents some techniques for selecting linguistically adequate hypotheses of new grammatical knowledge to be used as resources of grammatical knowledge acquisition. In our framework of linguistic knowledge acquisition, a rule-based hypothesis generator is invoked in case of parsing failures and all the possible hypotheses of new grammar rules or lexical entries are generated from partial parsing results. Although each hypothesis could recover the defects of the existing grammar, the greater part of hypotheses are linguistically unnatural. The techniques we propose here prevent such unnatural hypotheses from being generated without discarding plausible ones and make the following corpus-based acquisition process more efficient and more reliable.

1 Introduction

Reusability of existing linguistic knowledge is the most important requirement for the rapid development of practical natural language processing systems. In order to realize automatic customization of existing linguistic knowledge to each application domain, we proposed a new approach of linguistic knowledge acquisition, which is a combination of symbolic and statistical approaches [Kiyono and Tsujii, 1993].

The framework of our approach is shown in Figure 1. The acquisition flow starts with executing the parse of each sentence in a corpus. If parsing failed, the 'Hypothesis Generator' produces the hypotheses of additional grammatical knowledge, each of which could recover the incompleteness of the existing grammar. After iterating this hypothesis generation process for all the sentences in the corpus, the hypotheses are passed to the statistical analysis process and finally plausible hypotheses are chosen as new knowledge by observing statistical properties of the hypotheses.

Unlike robust parsing [Mellish, 1989; Goeser, 1992; Douglas and Dale, 1992] or non-statistical approach for grammar acquisition, our approach does not require a mechanism to detect the cause of the parsing failure in the sentencial analysis phase and therefore the 'Hypothesis Generator' may output all the possible hypotheses. However, the greater part of hypotheses generated by a simple deductive mechanism are unnatural revisions of the existing grammar. For example, even a rule which derives a top node category S directly from the input string of words might be hypothesized.

*also a staff member of Matsushita Electric Industrial Co.,Ltd., Shinagawa, Tokyo, JAPAN.

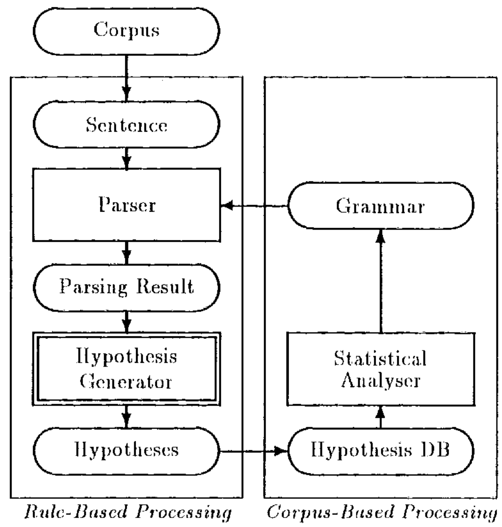

Figure 1: Framework of Grammar Acquisition

Linguistically unnatural hypotheses have harmful effects on the following corpus-based process, not only making the process inefficient but also interfering with statistical data as noise. In this paper, some techniques to remove such inadequate hypotheses are proposed and the results of experiments which show the effectiveness of the proposed techniques are also discussed.

2 Grammar Hypothesizing

2.1 Grammar Formalism

The grammar formalism we use is a conventional unification-based grammar. Each grammar rule is written in the form of a combination of a context-free rule and feature unification functions. This formalism is not specific to any linguistic theory, but we introduced a number of concepts widely accepted in linguistic theories, such as grammatical functions, subcategorization frames, and X-bar theory.

The parsing system we introduced to apply our grammar formalism is a system called SAX [Matsumoto, 1986]. SAX uses the concepts of active and inactive edges of *Chart Parsing* and analyses an input sentence with a bottom-up and parallel algorithm. As the grammar hypothesizing algorithm is supposed to refer partial parsing results of unsuccessfully parsed sentences, we slightly modified SAX so that it outputs inactive edges as partial parsing results.

2.2 Basic Algorithm

When SAX fails to parse a sentence, no inactive edge of category S spanning the whole sentence exists in the parsing result. Grammar hypothesizing is a process to introduce this inactive edge by augmenting the current grammar. The basic part of the hypothesis generation algorithm is written as follows:

[Algorithm] An inactive edge $[ie(A) : x_0, x_n]$ can be introduced from x_0 to x_n, with label A, by each of the hypotheses generated by the following two steps.

[Step 1] For each sequence of inactive edges, $[ie(B_1) : x_0, x_1], \ldots, [ie(B_n) : x_{n-1}, x_n]$, spanning from x_0 to x_n, generates a new rule.

$$A \Rightarrow B_1, \cdots, B_n$$

[Step 2] For each existing rule $A \Rightarrow A_1, \cdots, A_n$, find an incomplete sequence of inactive edges, $[ie(A_1) : x_0, x_1], \ldots, [ie(A_{i-1}) : x_{i-2}, x_{i-1}]$, $[ie(A_{i+1}) : x_i, x_{i+1}], \ldots, [ie(A_n) : x_{n-1}, x_n]$, and call this algorithm for $[ie(A_i) : x_{i-1}, x_i]$.

Feature Structures: A rule generated in **[Step 1]** could be a lexical entry when this top-down algorithm reaches the bottom. As we adopted a unification-based grammar formalism, we extended the algorithm so that it can hypothesize a feature structure of a lexical entry by observing surrounding successful categories. As the algorithm works even for a complex feature like a subcategorization frame, it can be used to acquire a subcategorization dictionary. While some previous works on subcategorization frame acquisition assumed very little prior knowledge concerning the classification of subcategorization frames [Brent, 1991; Manning, 1993], our approach assumes the existence of grammar rules specifying subcategorization frame assignment, which enables more accurate learning of subcategorization frames.

Multiple Defects: In **[Step 2]** of the algorithm, it is supposed that each unsuccessfully parsed sentence has exactly one cause of failure but a sentence in actual texts often contains two or more causes of failure (for example, two unknown words). To solve this problem, we extended the algorithm so that it searches for a multiple hypothesis which is a set of rewriting rules and lexical entries.

3 Hypothesis Selection

3.1 Basic Grammatical Constraints

From a linguistic point of view, hypotheses generated by the algorithm given above might contain many unnatural hypotheses because the algorithm itself does not have any linguistic knowledge to judge the appropriateness of hypotheses. To remove unnatural hypotheses, we have introduced the following criteria [Kiyono and Tsujii, 1993].

- The maximum number of adjacent unsuccessful categories is set to 2 in order not to decrease the efficiency of the algorithm.
- The maximum number of daughter nodes is set to 3.
- Supposing that the existing grammar contains all the category conversion rules, a unary rule which has only one daughter node is not generated.
- Using generalizations embodied in the existing grammar, a hypothesis containing a sequence of subnodes which are collected into a larger category by existing grammar rules is not generated.
- Distinguishing non-lexical categories from lexical categories, a hypothesis whose mother category is a lexical category is not generated.
- Assuming that the existing grammar has a complete set of functional words, a lexical hypothesis is restricted to the open lexical categories, such as noun, verb, adjective, and adverb.

3.2 Constraint based on Local Boundaries

A new constraint on the violation of the boundary condition given to phrases was introduced to avoid any collection of adjacent successful categories in rule hypothesizing. The boundary condition is given by putting parentheses at both ends of a phrase, such as a noun phrase, a verb phrase, and a prepositional phrase. This constraint filters out a hypothesis which crosses either end, not both ends, of a phrase. For example, when parentheses are put like "[The default blocking factor] is [20 blocks]", a hypothesis '$VP \Rightarrow VP, NP, VERBBE$' covering "blocking factor is" is discarded because of the violation of the boundary condition of a noun phrase "The default blocking factor".

This constraint requires the human task of putting parentheses before the hypothesis generator is invoked. In comparison with writing a constituent structure of the whole sentence, this work is much easier because we have only to give parentheses to definite phrases. Moreover, instead of giving parentheses by hand, we can even obtain various tagged corpora.

As this constraint is also applicable to other constituents of the input sentence, it might improve the efficiency of the top-down hypothesizing algorithm.

3.3 Constraint based on X-bar Theory

Most of the criteria in 3.1 are based on linguistic category classification but none of them commits itself to dealing with the relationship among the mother node and the daughter nodes. For example, supposing the existing grammar does not contain a rule for participial adjuncts in noun phrases, the hypothesizing program generates a new rewriting rule '$NP \Rightarrow VP, NP$' from the phrase "blocking factor" in the sentence "The default blocking factor is 20 blocks". However, the program also generates other alternative hypotheses from the same phrase, such as '$PP \Rightarrow VP, NP$', '$INFINITIVE \Rightarrow VP, NP$', and '$THAT\_CLAUSE \Rightarrow VP, NP$', each of which derives a post-positional adjunct for "default" by believing "default" is a head noun of the noun phrase. Linguistically, such combinations of mother nodes and daughter nodes are not allowed.

As a general principle for explaining phrase structures, *X-bar theory* is widely accepted. According to X-bar theory, a grammar rule is (or can be converted to) either of the following forms, where each prime(') expresses the projection level of a head X. The projection level increases as grammar rules are applied and X'' is called a *maximal projection* of that category. U and W are adjuncts of X' and should be maximal projections of some categories.

$X'' \Rightarrow Y X' Z$

$X' \Rightarrow U X W$

If the existing grammar is written in X-bar theory, this constraint is drastically effective in reducing the number of hypotheses.

3.4 Plausibility of Hypotheses

Among the hypotheses which passed through all the constraints, each one has a different plausibility as grammatical knowledge. Assuming that the existing grammar is reasonably comprehensive, lexical or idiosyncratic knowledge should be more plausible than general rewriting rules. In order to emphasize this tendency, each hypothesis is given the following plausibility value.

$$P(Hypo_i) = 1 - \frac{W(Hypo_i) \times H(Hypo_i)}{W(S) \times H(S)}$$

This value is related to the proportion of the size, or the product of the width and the height, of the subtree composed by the hypothesis in the whole structure of the sentence. The value ranges from 0 to 1 and gets bigger if the hypothesis covers a smaller part of the sentence. The width of the hypothesis, $W(Hypo_i)$, is defined as the word count of the subtree and the height $H(Hypo_i)$ is as the shortest path from lexical nodes to the top node of the subtree.

4 Experiments

4.1 Corpus

In order to check the effects of the hypothesis selection techniques, we carried out some experiments with the UNIX on-line manual. 100 sentences were chosen as an experimental set from the manual. The characteristics of this corpus are as follows.

- Number of sentences: 100
- Length of sentences: 9.08 words (average)
- Number of different words: 381
- Examples:

  There is no escape sequence that prints a double-quote.
  Use the next argument as the blocking factor for tape records.
  The default blocking factor is 20 blocks.
  ...

4.2 Given Grammatical Knowledge

Two sets of grammar rules were prepared for the experiments, *Grammar A* and *Grammar B*. Grammar A contains 118 rewriting rules that cover basic expressions of English. Grammar B is a subset of Grammar A and contains only 25 rewriting rules. The contents of Grammar A and Grammar B are shown in Table 1.

The dictionary we use is the *EDR English Dictionary* containing 200,000 entries. The entries of this dictionary are not written in the form of a feature structure but have the encoded information of the syntactic patterns, which we interpret as a feature structure. As the EDR Dictionary was developed as a master dictionary for various applications, it took in the information concerning all the appearances of each word without screening by frequencies. This characteristic of the EDR Dictionary increases the ambiguity of parsing. In fact, each word within the sample sentences from the UNIX manual has 1.49 parts of speech in the EDR Dictionary while the same value is 1.41 according to the *COLLINS COBUILD Dictionary*.

Mother Category	Grammar A	Grammar B
Sentence	23	1
Verb Phrase	40	12
Noun Phrase	27	7
Prepositional Phrase	2	1
Adjective Phrase	9	1
Adverbial Phrase	5	1
Infinitive Clause	4	1
That Clause	1	1
Relative Clause	6	0
Subordinate Clause	1	0
Total	118	25

Table 1: Rule Counts of Two Grammar Sets

4.3 Generated Hypotheses

General Outcome: The experiments of generating hypotheses were carried out with Grammar A under three different conditions, (a) using the basic grammatical constraints only, (b) adding the constraint with local phrasal boundaries given as parentheses, and (c) adding the constraint with X-bar theory. To carry out experiments (b) and (c), within the target sentences, parentheses were given to noun phrases, infinitive clauses, *that*-clauses, and subordinate clauses. A part of the result of experiment (a) is shown in Table 2, each column of which displays the number of hypotheses generated. The columns 'Single' and 'Multiple' show the numbers of single and multiple hypotheses respectively.

The results of the three experiments are summarized in Table 3. The parser failed to analyse 61 out of 100 sentences and the grammar hypothesizing program was invoked for those sentences. While no hypotheses were generated from 20 or 30% of unsuccessfully parsed sentences because the current hypothesizing algorithm does not allow vertical duplication of incompleteness and also because the parameters of the basic grammatical constraints do not allow the existence of more than two adjacent incomplete nodes, the results on the numbers of actual hypotheses made show that the stronger the constraint we pose, the fewer hypotheses are generated. The average hypotheses per sentence, calculated by dividing the total hypothesis count of 1,301 in (a), 708 in (b), and 231 in (c), by the number of actual sentences from which hypotheses were generated, 50 in (a), 44 in (b), and 41 in (c), was reduced from 26.0 to 5.6.

In some cases, all the hypotheses are removed by newly introduced constraints, 6 sentences by the local boundary constraint and 3 more sentences by the constraint of X-bar theory. Investigation of the initial set of hypotheses generated from such sentences revealed that no plausible hypothesis was included in it. Therefore, these sentences are not critical to the hypothesis selection method we introduced.

In the final set of hypotheses, 30 plausible hypothe-

Sentence	Single		Multiple			Total
	Lex	Rule	Lex	Mixed	Rule	
The default blocking factor is 20 blocks.	3	18	0	0	0	21
The output device in use is not capable of backspacing.	4	26	0	0	0	30
Remove initial definitions for all predefined symbols.	3	24	0	0	0	27
The escaped NEWLINE is not included in the macro value.	0	0	2	2	0	4
Components of an expression are separated by white space.	2	16	0	0	0	18
The name of this directory is listed in the folder variable.	3	0	0	0	0	3
The name of the editor is listed in the EDITOR variable.	2	0	0	0	0	2

Table 2: Part of the Result of Experiment (a)

	Experiment (a)	Experiment (b)	Experiment (c)
No. of Unsuccessfully Parsed Sentences	61	61	61
No. of Sentences which generated No Hypothesis	11	17	20
No. of Sentences which generated Single Hypotheses	43	39	37
No. of Sentences which generated Multiple Hypotheses	7	5	4
No. of Sentences which generated Plausible Hypotheses	33	32	30
Rank of Plausible Hypotheses (Average)	7.4	2.8	1.6
No. of Hypotheses (Total)	1301	708	231
No. of Hypotheses (Average)	26.0	16.1	5.6

Table 3: Hypotheses Generated from Different Conditions

ses, 7 new rewriting rules and 23 new or modified lexical entries, remained without being filtered out by newly introduced constraints. Some of the plausible hypotheses are listed below.

New Rule: np ⇒ np,adjp.
New Rule: np ⇒ np,np.
New Rule: np ⇒ vp,np. (from 3 sentences)
New Rule: np ⇒ vppsv,np.
New Rule: vp ⇒ vp,p.
New Lexical Entry: n ⇒ ['DELETE'].
New Lexical Entry: n ⇒ [pathnames].
Modified Lexical Entry: v ⇒ [default].
Modified Lexical Entry: adj ⇒ [invisible].
Modified Lexical Entry: adj ⇒ [capable].
New Lexical Entry: adv ⇒ [recursively].
......

The weighting function explained in 3.4 was not used for selecting hypotheses but the validity of it was proved by counting the order of each plausible hypothesis in the set of generated hypotheses. The row of 'Rank of Plausible Hypotheses' in Table 3 indicates that plausible hypotheses stand much higher than the middle of the order.

Examples: Hereafter, in order to show how hypotheses were selected by each constraint, we explain the results for some typical examples.

Ex.1) "The default blocking factor is 20 blocks."

As Grammar A does not contain a rule for participial adjuncts, the parser fails to analyse the noun phrase "the default blocking factor" and the grammar hypothesizing program is invoked. While this program generates 21 hypotheses in experiment (a), it filters out the following 12 hypotheses in experiment (b). While checking local boundary violation, the program removes those grammatically unnatural combinations of categories, though it does not use any linguistic knowledge.

```
New Rule:  advp => np,vp.
New Rule:  infinitive => np,vp.
New Rule:  infinitive => vp,np,vp.
New Rule:  np => s,vp,np.
New Rule:  np => vp,np,vp.
New Rule:  pp => np,vp.
New Rule:  pp => vp,np,vp.
New Rule:  that_clause => vp,np,vp.
New Rule:  vp => vp,np,auxbe.
New Rule:  vp => vp,np,s.
New Rule:  vppsv => vp,np,auxbe.
New Rule:  vppsv => vp,np,vp.
```

Moreover, the program filters out the following 4 hypothese with the constraint of X-bar theory.

```
New Rule:  infinitive => vp,np.
New Rule:  pp => vp,np.
New Rule:  that_clause => vp,np.
New Rule:  vppsv => vp,np.
```

Finally, the following 5 hypotheses, among which the expected hypothesis '$NP \Rightarrow VP, NP$' still remains, are generated.

```
Modified Lexical Entry:  n => [factor].
New Lexical Entry:  adv => [factor].
New Lexical Entry:  n => [blocking].
New Rule:  np => vp,np.
New Rule:  np => s,vp,np.
```

Ex.2) "The output device in use is not capable of backspacing."

This sentence is also parsed unsuccessfully because the current version of the EDR Dictionary does not have information that "capable" subcategorizes a prepositional phrase. Among the initial set of 30 hypotheses, the following 8 hypotheses pass through

	Grammar A	Grammar B
No. of Unsuccessfully Parsed Sentences	61	97
No. of Sentences which generated No Hypothesis	11	45
No. of Sentences which generated Single Hypotheses	43	41
No. of Sentences which generated Multiple Hypotheses	7	11
No. of Sentences which generated Plausible Hypotheses	31	16
No. of Hypotheses (Total)	1301	550
No. of Hypotheses (Average)	26.0	10.6

Table 4: Hypotheses Generated from Two Grammar Sets

the constraints of local boundaries and X-bar theory. The first hypothesis in the list is the plausible hypothesis obtained in search of the real cause of the feature disagreement between "capable" and "of backspacing". This lexical hypothesis for "capable" contains a modified version of its subcategorization frame so that it subcategorizes *of*-prepositional phrase.

```
Modified Lexical Entry:  adj => [capable].
New Lexical Entry:  n => [capable].
New Lexical Entry:  v => [capable].
New Lexical Entry:  v => [not].
New Rule:  adjp => neg,adjp.
New Rule:  adjp => neg,adjp.
New Rule:  s => s,adjp.
New Rule:  vp => vp,p.
```

4.4 Hypotheses from Smaller Knowledge

Another experiment was performed with Grammar B under the basic grammatical constraints in order to compare the effects of the maturity of existing grammatical knowledge. The numbers of hypotheses generated from two grammar sets are shown in Table 4.

The coverage of Grammar B is so limited that 97 out of 100 sentences were parsed unsuccessfully and passed to the Hypothesis Generator. However, as the immaturity of Grammar B also affects the number of generated hypotheses, the number of plausible hypotheses among the 550 hypotheses (10.6 hypotheses per sentence) generated from 97 sentences was only 16. This result claims that cyclic acquisition of grammatical knowledge is valid. Even the sentences from which no hypotheses are generated with a small grammar would be taken into consideration in a later acquisition cycle with a larger grammar.

5 Conclusion

This paper proposed techniques for selecting appropriate hypotheses in the rule-based processing stage of grammar acquisition. The experiments to examine the effects of these techniques indicate that they have several advantages.

- The newly introduced constraints reduce the number of hypotheses per sentence, from 26.0 to only 5.6, small enough to be treated in a corpus-based processing environment. This hypothesis selection is done without discarding plausible hypotheses. Although, all the initial hypotheses may be, in certain cases, removed by the new constraints, this happens only if no plausible hypothesis is included in the initial set.
- Even if no hypothesis is generated from an unsuccessfully parsed sentence (20 out of 61 sentences in experiment (c)) or no plausible hypothesis is included in the initial hypothesis set (11 out of 41 sentences in experiment (c)), a plausible hypothesis will be generated in the later acquisition cycle after adding grammatical knowledge vital for the sentence.
- Among the generated hypotheses, lexical hypotheses are more plausible than rule hypotheses (23 out of 30 plausible hypotheses were lexical in experiment (c)). This fact means that the grammar used for the experiments has an almost sufficient set of rewriting rules and that, after the grammar reaches such a mature situation during the acquisition cycle, only lexical or idiosyncratic knowledge has to be added. As our method has a facility to hypothesize a lexical entry with its feature structure including a subcategorization frame, we can set the target of acquisition only to lexical knowledge for a large dictionary.
- The local boundary constraint was introduced for automatic hypothesis selection, but it might also be used in an interactive debugging tool for grammar maintenance.

References

[Brent, 1991] Michael R. Brent. Automatic Acquisition of Subcategorization Frames from Untagged Text. In *Proc. of the 29st ACL meeting*, pp.209–214, 1991.

[Douglas and Dale, 1992] Shona Douglas and Robert Dale. Towards Robust PATR. In *Proc. of COLING-92*, pp.468–474, 1992.

[Goeser, 1992] Sebastian Goeser. Chart Parsing of Robust Grammars. In *Proc. of COLING-92*, pp.120–126, 1992.

[Kiyono and Tsujii, 1993] Masaki Kiyono and Jun'ichi Tsujii. Linguistic knowledge acquisition from parsing failures. In *Proc. of EACL-93*, pp.222–231, 1993.

[Manning, 1993] Christopher D. Manning. Automatic Acquisition of a Large Subcategorization Dictionary from Corpora. In *Proc. of the 31st ACL meeting*, pp.235–242, 1993.

[Matsumoto, 1986] Yuuji Matsumoto. A Parallel Parsing System for Natural Language Analysis. In *Lecture Notes in Computer Science 225*, Springer-Verlag, pp.394–409, 1986.

[Mellish, 1989] Chris S. Mellish. Some Chart-based Techniques for Parsing Ill-formed Input. In *Proc. of the 27th ACL meeting*, pp.102–109, 1989.