2A-4

Thai Syntax Analysis Using GPSG

(ＧＰＳＧを利用したタイ語の構文解析)

Vises Vorasucha and Hozumi Tanaka
Department of Computer Science, Tokyo Institute of Technology

Thai syntactic rules can be written very compactly in ID and LP rule format, which is a basic concept of GPSG. Using translator TIL, a mixed program of Prolog and DCG, we translate those syntactic rules into DCG format, and use the translated rules to do Thai syntax analysis.

1. Introduction

GPSG (Generalized Phrase Structure Grammar), which is a type of generative grammar, was mainly introduced by Gerald Gazdar and Geoffrey Pullum in 1982[1,2]. Most of Gazdar's works based on English, but GPSG itself can be applied to other languages too.

We tried to do syntax analysis of Thai, and found that GPSG can treat Thai in a very simple way by using ID and LP rule format, which is the basic concept of GPSG, to write Thai syntactic rules. We built a simple GPSG translator which translates rules that written in ID and LP rule format into DCG format. We call this translator TIL (Translator for ID and LP rules).

In this paper, we introduce some characteristics of Thai in Section 2, why Thai syntactic rules are easy to be written in ID/LP rule format. Besides standard format of ID/LP rule some special formats were added to TIL. We will discuss about these in Section 3, and show how we use TIL's ID/LP rules format to represent Thai syntactic rules in Section 4. In Section 5, we discuss about some topics on TIL. How to treat the augmentation in ID rules, how to apply other concept of GPSG to TIL. Finally, in Section 6, the Conclusion, we indicate the current state of our research, and our further work.

2. Some Characteristics of Thai

Thai sentences are similar to English sentences in the sense that they are beginning with subject, and followed with verb, and object. The differences occur with adjectives and adverbs. For example, in English, adjectives are placed before words they modify, but in Thai they are placed after the words. Normally, Thai sentences look like pattern in (1).

(1) S --> subj (adj) verb (obj (adj)) (adv).

In fact, according to Thai linguists, complete affirmative single sentences without embedded sentence have only 11 patterns as shown in (2).

(2)

s1 --> n.	s1 --> subj vp.
s1 --> n n.	s1 --> vp objd obji.
s1 --> vp.	s1 --> subj vp objd obji.
s1 --> vp objd.	s1 --> obji subj vp objd.
s1 --> subj vp objd.	s1 --> objd subj vp obji.
s1 --> objd subj vp.	

Let's take an attention at the last three patterns on the right hand side. All of them have the same non-terminal symbols, but the only difference is that they stand in different places. That is, in Thai non-terminal symbols are, although limitedly, movable in the sentence. Moreover, if we think that some of the non-terminal symbols are omissible, we can see that most of the rules are look alike. So they can be easily written in ID/LP rule format. This happens to other phrases in Thai, too. (3) is an example of noun phrases, which shows that they are made from less than four kinds of non-terminal symbols.

(3)

np --> mn.	np --> mn num.
np --> mn sbv num.	np --> mn sbv.
np --> mn num sbv.	np --> mn df.

```
np --> mn df sbv.          np --> mn sbv num df.
np --> mn df num.          np --> mn num sbv df.
np --> mn num df.          np --> mn sbv df num.
```

Especially for preposition phrases which modify time, place or other preposition phrases, they are really freely movable to any position in a sentence. We can use only one ID rule as shown in (4) with no LP rule, to represent about 50 patterns of these kinds of sentences. From this point of view, Thai syntactic rules are really suitable to be written in ID/LP rule format.

(4)
```
s --> (pp), (pt), (ps), s1.
```

We have explained why ID/LP rule format is suitable to Thai syntactic rules. Next we are going to explain about some characteristics of TIL.

3. Some Characteristics of TIL

At this moment, TIL can support rule systems which are written in ID/LP rule format with some simple features. The normal format of ID/LP rule for TIL look like this.

(5)
```
ID rule:                LP rule:
a --> b, c, d.          c << d.
b --> a, c, d.
```

In addition, we can define our own variables in any non-terminal symbols, and add Prolog program: an augmentation, to any rules, too. Non-terminal symbols without any variables defined in will be attached one variable by TIL automatically, for the sake of parser used in syntax analysis. For augmentation, this is done by enclosing a Prolog program with a left brace '{' and a right brace '}'. Presently, things enclosed between braces and the braces itself will be put at the end of the rule inevitably. By using brackets, TIL also can treat the omissible format of ID rules.

Therefore, TIL will translates ID/LP rules in (6) as shown.

(6)
```
ID Rule:                                LP Rule:
a --> b(Inf), c(Inf), d.                c << d.
b --> a(Inf), { Inf==[3] }, (c, d).

Translated Rule:
a(Ad) --> b(Inf), c(Inf), d(Ad3).
a(Ad) --> c(Inf), b(Inf), d(Ad3).
a(Ad) --> c(Inf), d(Ad2), b(Inf).
b(Ad) --> a(Inf), { Inf==[3] }.
b(Ad) --> a(Inf), c(Ad2), d(Ad3), { Inf==[3] }.
b(Ad) --> c(Ad1), a(Inf), d(Ad3), { Inf==[3] }.
b(Ad) --> c(Ad1), d(Ad2), a(Inf), { Inf==[3] }.
```

Augmentations may need to be put at a specified placed, like between preposition phrase and noun phrase, not at the end depend on the rules. We will discuss about this in Section 5.

For LP rule, we can add a list structured constant to each node as an argument, too. TIL can distinguish the difference between 2 nodes which have different constant, nevertheless they have the same node name, or not. LP rule will look like this.

(7)
```
subj << vp([3]) << vp([4]) << objd << obji.
```

Besides ID and LP rule, TIL lets user to be able to define grammar rule directly. TIL will skip those rules that begin with mark '+', and add them to the translated rules without changing them. (8) shows the format for this kind of rules.

(8)
```
+ bgst(Adb) --> obji(Ado1), subj(Ads2),
                vp(Adv), objd(Ado4),
                { Adv == [4] }.
```

We have mentioned about some special characteristics of TIL. Besides, standard GPSG's ID/LP rule format, we added some technics to TIL to let it more useful. Next we are going to show you how we use TIL to translate Thai syntactic rules.

4. ID and LP Rules for Thai

Some parts of Thai language grammar have been written in DCG form by Thai linguists already. Those DCG rules are almost for sentences which have only one verb. Here we reformed those rules into ID/LP rule format, and used TIL to translate it back to DCG format again to show that it is more simple to write Thai syntactic rules in ID/LP rule format. We omit to write all of the rules here, but will show only some interesting ones.

We have rules for noun phrases as shown in (9).

(9)
```
np --> mn, sbv.                          np --> mn(Adm), num(Adm).
np --> mn, df.                           np --> mn(Adm), sbv, num(Adm).
np --> mn, df, sbv.                      np --> mn(Adm), num(Adm), sbv.
np --> mn(Adm), df, num(Adm).            np --> mn(Adm), num(Adm), df.
np --> mn(Adm), sbv, num(Adm), df.       np --> mn(Adm), num(Adm), sbv, df.
np --> mn(Adm), sbv, df, num(Adm).
```

For these rules, we use ID and LP rules shown in (10) to represent them.

(10)
```
ID Rule:                                 LP Rule:
np --> mn(Adm), (sbv), (num(Adm)).       mn << sbv.
np --> mn(Adm), df, (num(Adm)).          mn << num.
 + np(Adnp) --> mn(Admn1), df(Addf2),    mn << df.
                sbv(Adsbv3).             sbv << df.
np --> mn(Adm), sbv, num(Adm), df.
```

We have rules for sentences that have single-object verb, like eat, hit, see and etc. shown in (11).

(11)
```
bgst --> subj, vp([3]), objd.            bgst --> vp([3]), objd.
bgst --> objd, subj ,vp([3]).
```

For these rules, we use ID and LP rules shown in (12) to represent them.

(12)
```
ID Rule:
bgst --> (subj),vp(Adv),objd, { Adv == [3] }.
+ bgst(Adbgst) --> objd(Adobjd1),
                   subj(Adsubj2),
                   vp(Adv),
                   { Adv == [3] }.
LP Rule:
subj << vp << objd.
```

Next are rules for sentences that have double-object verb, like teach, tell, give and etc. shown in (13), and (14) are ID and LP rules for them.

(13)
```
bgst --> vp([4]), objd, obji.            bgst --> subj, vp([4]), objd, obji.
bgst --> obji, subj, vp([4]), objd.      bgst --> objd, subj, vp([4]), obji.
```

(14)
```
ID Rule:
bgst --> (subj),vp(Adv),objd,obji, { Adv == [4] }.
 + bgst(Adbgst) --> obji(Adobji1), subj(Adsubj2),
                    vp(Adv), objd(Adobjd4),
                    { Adv == [4] }.
 + bgst(Adbgst) --> objd(Adobjd1), subj(Adsubj2),
                    vp(Adv), obji(Adobji4),
                    { Adv == [4] }.
LP Rule:
```

subj << vp << objd << obji.

Treating Thai syntactic rules in this way, we can largely reduce the number of rules which we have to maintain. Presently, we use 34 ID rules and 16 LP rules to represent 113 DCG rules.

5. Topics

At this moment, we have not applied any feature conventions to TIL. In fact, our syntactic categories have no internal structure at all. To define the structure, it requires more study of Thai to check that what kinds of information are needed in doing syntax analysis, and to be put into the structure of syntactic categories. However, we have prepared one variable in each non-terminal symbol so that feature conventions can be applied to TIL in future.

Also, we have not applied Metarule convention to TIL yet. The format for Metarule is not difficult, but the problem is that it is still not clear whether we should applied Metarule to syntactic rules before parsing or during parsing.

One of TIL formats that different from standard GPSG formats is an augmentation addable format, which is an important tool to help us to write syntactic rules more precisely and simply. Standard GPSG have provided many conventions, like foot feature convention, and head feature convention, for making constrain in syntactic rules. But, there may be some constrains that can be written very simply by Prolog rather than by GPSG's conventions. Also, there may be some lower level controls we want to make in a syntactic rule but cannot write by GPSG's convention. TIL provides augmentation for these occasions. Now augmentations are replaced to the end of rules inevitably. However, sometime it is possible that augmentations need to be put not at the end of rules, but between any specified syntactic phrases. We can do this by attaching to each augmentation in ID rules an identified number. Then, in LP rules, we can freely define the order of augmentations as if they were non-terminal symbols by using identified number. The format will look like (15).

(15)

ID rule:	LP rule:
a --> ... , {1:...}, ... , {2: ...}, ...	... << {1:} << {2:} << ... .

6. Conclusion

We have built a basic GPSG translator:TIL showed its special characteristics, and how it works. Also, we showed that, for Thai syntactic rules, in ID/LP format instead of DCG format, we can reduce more than half of rules which we have to maintain.

We are going to extend TIL's functions, make usage of feature convention, and introduce the metarule convention.

Acknowledgement

We wish to thank assistant professor Noboru Akiyama and Wilai Hongladaromp for giving us Thai character font program to support this research, to Satoshi Konno for helping us to rewrite the program until it can connect to the main computer. Also thanks to students in Tanaka Laboratory for many helpful comments on the research and this paper.

References

[1] Gazdar, Gerald, and Geoffrey Pullum (1982) *Generalized phrase structure grammar: a theoretical synopsis*. Mimeo, Indiana University Linguistics Club, August 1982.

[2] Gazdar, Gerald, Ewan Klein, Geoffrey K. Pullum, and Ivan A. Sag (1982) *Coordinate Structure and Unbounded Dependencies*. In M. Barlow, D. Flickinger & I.A. Sag (eds.) Developments in Generalized Phrase Structure Grammar: Stanford Working Papers in Grammatical Theory, Volume 2. Bloomington: Indiana University Linguistics Clubs, 38-68, November 1982.

[3] Ngamsut, Chinda (1981) ภาษาศาสตร์ภาษาไทย : Odian Store, Bangkok, March 1981.

[4] Thichinpong, Preecha (1980) ศัพท์และภาษาไทย : Odian Store, Bangkok, November 1980.