A New Predictive Analyzer of English

Hiroayuki Musha

Department of Information Science
Tokyo Institute of Technology
Ohokayama, Meguro-ku, Tokyo 152, JAPAN

ABSTRACT

Aspects of syntactic predictions made during the recognition of English sentences are investigated. We reinforce Kuno’s original predictive analyzer[1] by introducing five types of predictions. For each type of prediction, we discuss and present its necessity, its description method, and recognition mechanism. We make use of three kinds of stacks whose behavior is specified by grammar rules in an extended version of Greibach normal form. We also investigate other factors that affect the predictive recognition process, i.e., preferences among syntactic ambiguities and necessary amount of lookahead. These factors as well as the proposed handling mechanisms of predictions are tested by analyzing two kinds of articles. In our experiment, more than seventy percent of sentences are recognized and looking two words ahead seems to be the critical length for the predictive recognition.

1. Introduction

When human reads normal sentences, we rarely feel something is wrong with the structure we are constructing and are seldom compelled to backtrack for reconstructing an alternative. If we could simulate the internal mechanism that makes it possible to select deterministically the unique syntactic structure in a simple way, we may be able to construct more natural and efficient language processing systems. In this paper, we focus our attention on syntax of natural languages, particularly English, and predictions or expectations that can be made solely with syntactic information during the sentence recognition process are analyzed in detail. It includes machine executable mechanisms that enable proper handling of analyzed aspects and a description method of the mechanisms as grammar rules. The recognition method can be seen as a deterministic one [2] if we permit looking some words ahead. Also included in this paper are results of an experimental analysis in which more than seventy percent of sentences are recognized.

An analyzer which gives special attention to predictions was once developed by Kuno [1]. The analyzer makes use of the simple stack mechanism whose behavior is specified by rules described in Greibach normal form. In the method, however, we can find several kinds of rules that do not correspond to human predictive recognition process, which will be pointed out in this paper.

The following discussion is based mainly on the author’s (subjective) retrospect of the recognition process of English sentences. The author’s mother tongue is Japanese and he has been learning English as a second language. It seems to the author that he can understand better how he recognizes English than how he recognizes Japanese since he has been learning English consciously and can observe rather objectively the process of recognition.

The rest of this paper is organized as follows. In the next section, we discuss aspects of predictions, laying stress upon their proper handling by computers. The following section presents the results of an experiment. The conclusions are presented in the last section.

2. Aspects of Predictions

While reading or hearing English, we constantly predict or expect what may follow next. Such predictions can be classified into six types which we will describe below.

2.1. Essential Predictions

The simplest type of prediction, which forms the basis of the following discussions, is presented in this subsection. The characteristic of this type of prediction is that it is essential in forming an acceptable sentence structure.

Phrase structure grammar rules, especially those in Greibach normal form, can naturally describe this kind of prediction: we can consider the terminal symbol (or the lexical category) on the right-hand side of a rule as the current word and the nonterminal symbols that follow the terminal symbol as new predictions [3]. For example, the following rule describes what we predict when we encounter a transitive verb at the beginning of a verb phrase.

\[ VP \rightarrow vt \ NP \]

Note that the new prediction, NP, is essential to form a verb phrase. By adopting this kind of rules as a means of structural description of sentences, we can easily capture the structures by using the stack mechanism [1].

In the following subsections, except for the last subsection, these rules and the mechanism are gradually reinforced in order to handle a newly introduced prediction type. The extended mechanism provides us with a simpler (yet still powerful) means for recognition of sentence structures than, for example, ATN framework [4]. Other factors that affect the predictive recognition process are discussed in the last subsection.

2.2. Optional Predictions

We now extend our recognition mechanism by introducing optional predictions. This type of prediction is needed to handle postpositional modifiers that are not essential to form a sentence.

In the previous subsection, we saw that rules in Greibach normal form are suitable for expressing our predictive recognition process, but any rule should not predict too much. Consider the following rule that explains a possible structure of noun phrases.

\[ NP \rightarrow \text{article} \ NP-\text{ART} \, \text{ADJ_CLAUSE} \]

Concerning the correspondence with human language understanding process, however, the rule cannot be considered a good simulation of our understanding process: we predict a postpositional modifier, like an adjectival clause, not at the beginning of a noun phrase but at the beginning of the modifier. For our purpose, therefore, we must exclude this kind of rule that do not express our predictions properly.

Optional predictions are used to capture these structures. Here, we also extend the rule description to keep the correspondence between the grammar rules and the recognizing mechanism: we introduce the shifting flag.

The following rules are used to capture postpositional modifiers.

\[ \text{CW} \, \text{CP} \, \text{SF} \, \text{NP} \]

(1) art NP \rightarrow t \, \text{NP-ART}
(2) noun NP-ART \rightarrow t \, *\text{NP-N}
(3) rel_pro NP-N \rightarrow nil \, \text{ADJ_CLAUSE}

The first rule, for example, can be interpreted as follows: IF the current word (CW) is an article and the current prediction (CP: the top element of the stack) is NP, THEN shift the current word pointer (since the shifting flag (SF) is t) and replace the current prediction by the new prediction (NP).

The shifting flag enables us to proceed two or more state changes while looking at a single word. By using these notations
and the rules we can specify the state changes of the stack as shown in Figure 2-1. The prediction NP-N, with a prefix '*' which shows it is optional, is interpreted as the state in which a noun essential to form a noun phrase has already appeared and it may end there. It will be popped out from the stack or will be replaced by a new prediction according to the word that follows.

\[
\begin{array}{c}
\text{NP} \\
\text{PERIOD}
\end{array} \rightarrow \begin{array}{c}
\text{NP-ART} \\
\text{PERIOD}
\end{array} \rightarrow \begin{array}{c}
\text{NP-N} \\
\text{PERIOD}
\end{array} \rightarrow \begin{array}{c}
\text{ADJ,CL} \\
\text{PERIOD}
\end{array} \rightarrow \begin{array}{c}
\text{NP} \\
\text{PERIOD}
\end{array}
\]

\[
\text{NP} \rightarrow \text{noun}
\]

Figure 2-1. Handling of the optional prediction.

2.3. Bunch Predictions

We extend our model by introducing bunch predictions which enable us to predict a set of syntactic categories simultaneously. In the following subsection, we see that this kind of prediction is useful for handling coordinate conjunctions, too.

Various kinds of syntactic units can follow the verb be in a verb phrase and we cannot selectively predict one of these possibilities when we are reading the words, such as am or were, etc. The bunch predictions we introduce enable us to cope with this kind of predictions.

The following rule shows how to write a bunch prediction in a rule.

\[(\text{be fut}) (\text{VP fut}) \rightarrow \begin{array}{c}
\text{bunch (NP) (ADJ) (VP ing))} \\
\text{VP_MOD}
\end{array}\]

When a bunch prediction is pushed onto the stack, it works as if it were a single prediction until it becomes the top of the stack, and one of the constituent of the bunch prediction is, then, chosen to be appropriate according to the word encountered.

2.4. And Stack

In this subsection, we introduce another stack called the and stack to handle coordinate conjunctions. The method described here resembles that in [5] or [6], but with the and stack we can handle them quite simply.

The appearance of coordinate conjunctions are usually not predictable and it triggers a new kind of operation. Let us consider the following sentences.

(1) Mary had a little lamb and a kitten.
(2) Mary had a little lamb and she was always with him.
(3) Mary had a little lamb and she was always with him.

Conventional phrase structure grammar rules like:

\[S \rightarrow S \text{ and } S\]

are not directly useful for predictive recognition of the sentences. The structure that follows and depends not on the word itself but on the proceeding syntactic units being constructed. In the above sentences, a noun phrase, a verb phrase, and a clause are being constructed before the word, and each of these categories reappear in each of the three sentences, respectively.

By using the and stack, we can easily recognize these structures. Figure 2-2 shows the relationship between the prediction stack (the stack that holds predictions) and the and stack where unnecessary details are omitted. At stage (ii), the first prediction is replaced by two predictions NP and VP only by looking the first word Mary. The lower element of the and stack is changed to VP (S), which shows that while the VP of the prediction stack is being processed, we are constructing both VP and S. In the same way, the stacks change their states as shown in the figure and a list (NP VP S) is made and pushed on the and stack when we reach the word and. The only thing we have to do is that we make a bunch prediction [bunch (NP) (VP) (S)] and replace *NP-N by the bunch prediction. By looking at the words that follow we can choose one of the constituent predictions of the bunch prediction and process the rest of the sentence.

Figure 2-3 shows the handling of the optional prediction.
Finally found the old man and woman with the telescope.

Figure 3-1. Tree structure constructed by the analyzer.