Supporting Grammar Development Using Local Contextual Information

Yasunobu Kawaguchi

School of Information Science,
Japan Advanced Institute of Science and Technology

February 14, 1997

Keywords: grammar development, hypothesis selection, local contextual information.

In this thesis, we present a new method for supporting grammar developing process by utilizing local contextual information.

For practical use of natural language processing applications, we need a suitable and wide-coverage grammar. However, in general, constructing such grammar by hand is a time-consuming and hard task due to the complex phenonmena of a language. This motivates many researchers to study a method to automate a process of grammar construction. In recent years, there have been several attempts to learn grammars from scratch by using statistics from various formats of large hand-constructed corpora. However, the acquired grammars are, in some senses, rough and specific for some certain domains. Therefore, there is a need to study a method for customizing an existing grammar for practical applications. In this work, we focus on this point and propose a support method for grammar development where a grammar is refined by using local contextual information from a corpus.

Although there are at least three possible operations: "add a new rule" , "delete a rule" and "divide a category" in a general grammar development process, this work focuses on a method of adding a new rule in the first place. The aim is to construct a system which can give a user some plausible hypotheses which are candidates of the rule we should add to the original grammar.

The user interaction model we assume is simple: if a user selects a sentence which cannot be parsed by the current grammar, the system shows him some plausible hypotheses of the defect in the sentence, as well as candidates of a new rule.

The grammar formalism occupied in this work is a phrase structure context free grammar(PS–CFG). This formalism is popular and widely applied in many NLP applications due to its simplicity. The parser in our system is a simple bottom-up chart parser.

In our approach, there are two different types of grammar hypotheses called "Instance (Local) Hypothsis" and "Global Hypothesis". The former represents an inactive edge (this is a term used in the chart parsing methodology) which can cover a defect part of a sentence when it is introduced in parsing process. The latter represents a grammar rule which can produce inactive edges represented by instance hypotheses when it is introduced to the existing grammar.

The framework we adopted here is similar to the approach proposed Kiyono and Tsujii at 1994. In this framework, the grammar development process consists of two modules as follows.

- Hypothesis Generation
  This process is a rule-based approach. It generates a set of instance hypotheses which recover a defect of the existing grammar.

- Hypothesis Selection
  This process is a corpus-based statistical approach. It filters the generated instance hypotheses by using an environmental (local context) similarity measure.

In the Hypothesis Generation module, a set of possible instance hypotheses are generated by exploiting inactive edges (those of incomplete analysis of a sentence) and the current grammar. Each generated hypothesis can lead to complete the analysis of the sentence. However, in general, there maybe a large number of hypotheses generated and most of them are not always suitable that lead to incorrect analysis. After Hypothesis Generation which is a rule-based approach, we adopt a corpus-based approach to cut off useless and inappropriate hypotheses.

In the hypothesis selection task, we use "local contextual information" to induce and represent the plausibility of the hypotheses. The term "local contextual information" considered here is represented by a pair of categories immediately before and after the incomplete part of the sentence.

Our scoring method is to calculate and compare the local contextual information of an instance hypothesis with the same label in the correct analyzed sentences. For example, if an instance hypothesis $hypo(A)$(A is its category label) connects with category L on the left and R on the right, we estimate $hypo(A)$'s plausibility by how frequently the tuple, (L,A,R), appears in the whole corpus of correctly analyzed sentences. So, we call such tuple(a set of three categories), (L,A,R), as "category trigram".

In order to know how effective our scoring method works, we carried out an experiment with a grammar with 272 rules and nearly five thousand sentences which cannot be parsed by the grammar. The initial grammar is automatically learned from a bracketed corpus, the EDR English corpus, and this grammar can cover around 60% sentences in the corpus.

At first, we got a category trigram (such as L,A,R) from a corpus of successfully and correctly parsed sentences. The test corpus includes 5082 sentences which cannot be parsed by the current grammar.

Among these sentences, the hypothesis generation module could generate instance hypotheses for 4730 sentenses, but there were 1513 sentences which have no correct hypothesis. Therefore, 1865 sentences were regarded as failure of hypothesis generation. For the rest 3217 sentences, our system can find out correct hypotheses with many incorrect hypotheses. We examined the performance of our hypothesis selection on these 3217 sentences which include both correct and incorrect hypotheses. As a result, when we select only the hypothesis with the highest score as the correct hypothesis, the selected hypothesis was correct for 1362 sentences. If the system showed 10 highest hypotheses in the order of its scores, they contain some correct hypotheses for 2634 sentences(that is 81.9%).

In the experiment mentioned above, we used all of the categories as local contextual information. We also make an experiment in the case that only lexical categories are used as local contextual information. The result show the performance of the hypothesis selection of this case was a little worse.

This paper consists of five chapters. Chapter 1 gives an introduction for this research. In chapter 2, we describe some previous researchers and their hypothesis generation algorithm and hypothesis selection, and in the next chapter, the chapter 3, we describe our hypothesis scoreing method. In chapter 4, we describe the experiment and discuss the result. Finally, the last chapter gives some conclusions and some future works.