A Language Acquisition Model on a Computer using Partial Pattern Matching

Eigo Mori, Yoshikazu Miyanaga, Koji Tochinai
Department of Electronics and Information Engineering
Faculty of Engineering, Hokkaido University
N 13-Jo, W 8-Chome, Kita-ku, Sapporo, 060 Japan
Phone:(+81-11)706-6823 Fax:(+81-11)709-6277
{eigo,miyanaga,tochinai}@hudk.hokudai.ac.jp

Kenji Araki
Department of Electronics and Information Engineering
Faculty of Engineering, Hokkai-Gakuen University
S 26-Jo, W 11-Chome, Chuo-ku, Sapporo, 064 Japan
Phone:(+81-11)841-1161 ext.776 Fax:(81-11)551-2951
araki@eli.hokkai-s-u.ac.jp

We are interested in constructing a Language Acquisition Model on a computer, and intend to apply this learning model to Natural Language Processing (NLP). We propose here a method for acquiring translation rules between actual utterances and 'semantic expressions'.

The semantic expressions are intended to represent the state of mind of the hearer when he or she heard the utterances. We designed the program for acquiring a set of language rules by searching for matching patterns in examples for learning. This method is based on an idea that considers language as being constructed of a series of small rules superimposed on each other.

In our method, a given semantic expression only includes information specific to the corresponding utterance. This condition makes it easier to acquire rules, however it differs from the actual situation in the very early stages of human language acquisition. Therefore, at present, we view our system as a model of the later stages of acquisition.

In view of NLP, the biggest advantage of using the learning algorithm is that we do not need to worry about managing the rules. The system collects rules for itself automatically, and selects a suitable rule by considering the context. More specifically, it selects a rule to apply to a given utterance by calculating the similarity between semantic expressions derived from the utterance and its store of translation rules that correspond to these expressions. Similarity is calculated between semantic expressions (not between utterances) using a thesaurus, if it is needed, making it possible to select rules precisely by using features such as gender and person.

We constructed the actual system and confirmed its efficiency by experiment. The system was given about 600 pairs of utterances and semantic expressions, extracted from an elementary level English textbook, and produced about 100 rules. Acquired rules succeeded in translating 47% of 'unknown' utterances (also extracted from the textbook), and 99% of 'known' utterances.

The main reason for the failures in translation was because the system did not have enough information on each rule. This would be improved by exposing the system to more data. However, because the system checks all the rules it has, such an experiment would be infeasible because of the huge amount of time needed for processing. Therefore our next step is to design a faster rule access method, and we are now planning a method using word association lists.