International Cooperation for Building Large Scale Electronic Dictionary

TANAKA Hozumi

Tokyo Institute of Technology

International Cooperation for Building Large Scale Electronic Dictionary

Rescarches on natural language processing have been very active in Japan and the researchers who have engaged in the field have developed practically useful natural language understanding systems including machine translation systems. To make a current natural language understanding system more intelligent, we have to solve many difficult problems one of which is semantic processing. Many researchers will agree that the lack of knowledge in a (electronic) dictionary creates some of difficulties for semantic processing.

Semantic processing needs a lot of world knowledge as well as grammatical knowledge. Most of them should be encoded in a dictionary but we do not know exactly how to encode world knowledge. Developing a dictionary is not an easy task since the organization of knowledge in a dictionary relates directly to the problem of knowledge representation, which is one of the most difficult research themes in artificial intelligence.

The problem mentioned above is on the quality of a dictionary, but we have to solve a quantitative problem, namely the size of a dictionary. Suppose we are going to develop a dictionary, which has more than one hundred thousand entries, immense of manpower must be invested.

Because of the prematured level of research on knowledge representation, we cannot represent any kind of knowledge in our dictionary. Therefore the size of knowledge in each entry of a dictionary will be limited to be small. As mentioned before, however, it is necessary for us to make a lot of efforts to build a large scale dictionary which has more than one hundred thousand entries. We have to start building a large scale dictionary as soon as possible. At first the knowledge in each entry of the dictionary will not be enough but progressing our research, we can accumulate new knowledge on it. It is worth starting from where we can do.

In the past, dictionary development efforts have been made independently by many countries and by many research groups. They have developed their own dictionaries based on their own theories. As the results, many duplicated efforts have been made. To avoid the duplicated efforts, we would like to develop a unified view through which many dictionaries will be constructed. It is clear that such a view will not be obtained without joint efforts by international cooperations. With regard to a machine translation system, developing multilingual translation dictionaries is very

important. Without international cooperation, better multilingual dictionaries will not be produced.

To construct better large scale dictionaries, we need many international cooperations. What we have to discuss at first is to decide what kind of information (or knowledge) will be required for the natural language research. At first glance, there will be a lot of controversies about it but we will be able to get a consensus about the type of knowledge we want to have. As our joint research proceeds, the coverage of consensus will more and more enlarge.

The second step is to design a common language to describe a dictionary. Through our language designing process, it is possible for us to get many ideas about an interlingua. International cooperation will help us create our image of an interlingua. This will be an interesting research theme especially not only for describing dictionaries but also for developing multilingual machine translation systems.

The third step is to build an international research center in which many researchers from many countries will be worked together to develop many kinds of dictionaries which will be produced under a common and unified view. The center will take a role of maintenance of dictionaries and keep them the newest version by collecting all technical wards born in the world.

We would like to make a final comment on our dictionaries which will be developed through international cooperations. We hope that the dictionaries developed at the international research center should not be monopolized. In other words, we have to make every dictionary open to anyone who want to use it.