Automatic Discovery of Definition Patterns Based on the MDL Principle

Masatoshi Tsuchiya and Sadao Kurohashi

Graduate School of Informatics, Kyoto University
{tsuchiya,kuro}@pine.kuee.kyoto-u.ac.jp

1 Introduction

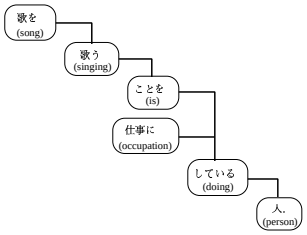
A number of approaches to construct a knowledge base for machines have been proposed. The Cyc project[1] manually constructed a knowledge base by using an artificial knowledge representation language. Such an approach, however, has not succeeded because of the heavy cost of construction and the difficulty in maintaining its consistency.

On the other hand, there is a growing interest in approaches which generate a knowledge base from a large amount of text corpora automatically. Among several types of corpora, dictionaries are a promising resource for a knowledge base[2, 3].

A dictionary consists of a set of definitional sentences about words or concepts, usually using sentential patterns. This paper proposes a method of discovering these definition patterns. Once definition patterns are discovered, they can be useful for automatically generating of knowledge base from a dictionary.

2 Discovery of Definition Patterns

The structure of a Japanese sentence can be described well by the dependency relation between *bunsetsu*s. A *bunsetsu* is a basic unit in the Japanese language, consisting of one or more content words and the following zero or more function words. The dependency structure of a Japanese sentence can be represented as a graph, in which *bunsetsu*s map to vertices, and dependency relations between *bunsetsu*s map to edges, as shown on the left.



In such a representation, any sub-sentential pattern can be regarded as a subgraph. The definition patterns which we want to discover automatically are fixed and important phrases or clauses. It is difficult to define exactly what the definition patterns are, however, they meet at least the following three criteria: 1) they probably occur frequently, 2) the bigger they are, the more important they seem to be, and 3) they can include semantic classes, instead of real words.

Since these criteria are trade-offs, we need an evaluation function to balance them. Therefore, we employ the Minimum Description Length (MDL) principle, proposed by Rissanen[4]. The MDL principle is a principle for both data compression and statistical estimation.

In our task, we define the description length of a set of graphs, each of which represents a definition sentence. Then, we look for a subgraph which minimizes the description length when the occurrences of the subgraph are reduced into a new single vertex. The detected subgraph is supposed to be a good definition pattern. Our method iterates the above process until such a subgraph cannot be detected anymore.

3 Algorithm

In order to reduce the size of the search space, we place two restrictions.

- Only the top-n frequent pairs of vertices are considered as the candidate subgraphs[1].
- Only the semantic classes that exist within the thesaurus in the form of a cut of a tree are considered[5]

4 Experiment and Discussion

In our experiment, Reikai Shogaku Kokugojiten, a Japanese dictionary for children, and Bunruigoihyou, a Japanese thesaurus, were used.

From this dictionary, we discovered many definition patterns which satisfy the three criteria mentioned in Sec. 2. For example, "*Shuto-wa* 'capital is' *PLACE*" (*PLACE* denotes a semantic class) was discovered from the definition sentences for *America*, *Japan*, and so on. "*Shigoto-ni* 'occupation' *shite-iru* 'doing' *hito* 'person' " was discovered from *singer*, *painter*, *cook* and others.

We are planning to investigate the detected definition patterns from the linguistic view point. We also have to improve our search algorithm.

References

1. Lenat, D. B. and Guha, R. V.: *Building Large Knowledge-Based Systems; Representation and Inference in the Cyc Project*, Addison Wesley Publishing Company, Inc. (1989).
2. Richardson, S. D., Dolan, W. B. and Vanderwende, L.: Mindnet : acquiring and structuring semantic information from text, *Proc. of COLING-ACL'98* (1998).
3. Kurohashi, S. and Sakai, Y.: Semantic Analysis of Japanese Noun Phrases : A New Approach to Dictionary-Based Understanding, *Proc. of ACL'99* (1999).
4. Rissanen, J.: *Stochastic Complexity in Stochastic Inquiry*, World Scientific Publishing Company (1989).
5. Li, H.: Generalizing Case Frames Using a Thesaurus and the MDL Principle, *Computational Linguistics*, Vol. 24, No. 2, pp. 217–244 (1998).

[1] The iterative detection of pairs of vertices can find bigger subgraphs consisting of three or more vertices. For example, A–B is first detected and reduced into A', then A'–C is detected, which means that we find A–B–C subgraph.