A Principle-Based Approach for Natural Language Processing

Eliminating the shortcomings of rule-based systems with statistics

-

- Wen-Lian Hsu
- Institute of Information Science, Academia Sinica

In natural language processing, an important task is to recognize various linguistic expressions. Many such expressions can be represented as rules or templates. These templates are matched by computer to identify those linguistic objects in text. However, in real world, there always seem to be many exceptions or variations not covered by rules or templates. A typical approach to cope with this situation is either to produce more templates or to relax the constraints of the templates (e.g., by inserting options or wild cards). But the former could create many similar case-by-case templates with no end in sight; and the latter could lead to lots of false positives, namely, matched but undesired linguistic expressions. Thus, the flexibility of rule matching has troubled the natural language processing (NLP) as well as the artificial intelligence (AI) community for years so as to make people believe that rule-based approach is not suitable for NLP or AI in general. On the other hand, fine-grained linguistic knowledge cannot be easily captured by current machine learning models, which resulted in mediocre recognition accuracy. Therefore, how to make the best out of rule-based and statistical approaches has been a very challenging task in natural language processing.

This paper describes a partial matching scheme that enables a single template to match a lot of semantically similar expressions with high accuracy, which we refer to as the Principle-Based Approach (PBA).

In PBA, we use a collection of frames to represent linguistic concepts or rules. Each frame is a collection of slots (also called components) with relations specified among them. A slot can be a word, phrase, semantic category, or another frame concept. One can specify position relations, collocation relations, and agreement relations and others among its slots. Unlike normal templates that involve mostly left-right relations among its components in a sentence, relations within frames can be multi-dimensional. For example, one slot could be a variable indicating the topic which other slots belong to.

To illustrate our partial matching scheme, consider a simple frame concept involving

5 components such that their relations in a sentence are arranged as 1, 2, 3, 4, 5 from left to right. Suppose in a sentence we can identify components 2, 3, and 5 in that order. So 1 and 4 are missing (deletion), and there maybe words inserting between 2 and 3 (insertion), and also between 3 and 5. Furthermore, a match for slot 5 could be on word-sense rather than on the word themselves (substitution). Our partial matching scheme allows for insertion, deletion and substitution. An insertion is given a positive score if it tends to collocate with its left or right matched components in general (otherwise, negative). A deletion can be harmless if slots 2, 3, and 5 contain a key combination for the frame. Note that many such key combinations can be pre-specified as indices of the frame. Collocation and bigram statistics can be incorporated in such score estimation. A substitution is given a lower score depending on their closeness in a semantic tree. After all these scores are determined, we can use an alignment algorithm to measure the fitness score and to decide how well the frame matches with the sentence.

PBA is inspired by the fact that when one studies a foreign language, he or she is usually presented with a collection of rules. These rules and their possible extensions and variations are practiced over and over again in real life to be mastered by the learner. PBA is flexible in that, it tends to relieve the burden of having to match with something “exactly” as specified and fine-grained linguistic knowledge can be more easily adopted to help estimate the scores of insertion, deletion and substitution in a PBA frame match.

We believe PBA can model more linguistic phenomena than current machine learning models, and is more suitable for NLP and AI in general. More details and examples of PBA will be covered in the talk.