An Investigative Study on Predicate-Argument Structure Analysis

Hiroyuki Yamagishi（0910904）

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

November 23, 2012

Keywords：Predicate-Argument structure, Case frame, Semantic role labeling, Machine learning.

In a sentence, a predicate takes a form like "X destroys the Y" and "X is beautiful" to describe the event the sentence represents, and nouns participate in the form inevitably. A structure that nouns center around a predicate represents the gist of the meaning of the sentence. If the structure is extracted properly, it will be very useful for a computer to treat the meaning of the sentence. Such a structure is called "predicate-argument Structure". A task that extracts predicate-argument structure from a sentences is called "predicate-argument structure analysis." This paper describes an investigative study predicate-argument structure analysis that is an important task in natural language processing (NLP).

In recent years, attention to the predicate argument structure analysis has been growing and studies on it became a shared task in the International Conference on Computational Natural Language Learning (CoNLL). Accordingly, a wide range and a number of research results have been reported. Predicate argument structure analysis is viewed as a task that tags words in a sentence with appropriate labels. Tagged labels are divided

roughly into surface and deep cases. A surface case is a grammatical role which an argument word plays in a sentence and a deep case is a semantic role. A surface and deep cases do not correspond one-to-one in general. As a deep case express semantic information, if we can tag words with deep case labels properly, we expect that we can improve the accuracy of question answering, machine translation, document summarization and information extraction.

The early studies used systems that tag sentences based on rules created manually. In 1998, FrameNet, a large scale corpus with semantic information was released. In 2002, Gildea and Jurafsky used FrameNet and reported automatic statistical analysis techniques based on machine learning. It was epoch-making in the history of this field. As the similar corpus, In 2005, Proposition Bank, which was made by tagging semantic role to Penn TreeBank corpus, was released. It promoted a paradigm shift from rule- based approach to statistics-based approach. In this trend, this task has been called "semantic role labeling," which was named by Gildea and Jurafsky, instead of "deep case analysis."

On the other hand, there are nouns called "Event-nouns," which represent situations. Gerber et al. reported a study on the predicate-argument structure analysis of event-nouns in 2010 and have attracted attention in recent years as an important topic.

Along such research history, this thesis describes an investigative study on major tasks in predicate-argument structure analysis, surface case labeling, semantic role labeling (deep case labeling) and event-nouns analysis.