JAPANESE SENTENCE ANALYSIS SYSTEM ESSAY - EVALUATION OF DICTIONARY DERIVED FROM REAL TEXT DATA

K. Shirai, J. Kubota, Y. Hayashi

Department of Electrical Engineering, WASEDA University

In this paper, we report on an experimental system of Japanese sentence analysis, called ESSAY.

Many Japanese sentence analysis systems, not only Phrase structure analysis systems but also Kakari-Uke analysis systems, are usually based on rules in syntactic level or Case grammatical restriction.

Comparing with such systems, our system is unique in the dictionary. In this dictionary, function of the language elements, such as words or auxiliary morpheme, are described. And these lexical entries are automatically constructed from analysis of real text data.

In order to evaluate the usefulness of such dictionary, we are accumulating Japanese sentences data, and applying statistical and structural analysis method to this data.

In the following we concentrate upon next 2 points.
(1) construction of dictionary
(2) overview of ESSAY

(1) Construction of dictionary

As the initial data we entered about 2000 sentences of elementary school text in Kana-letter (Japanese syllabary) not in Kanji (Chinese character).

Japanese is an agglutinative language, so in analyzing sentences they are usually separated into number of parts

called Bunsetsu. In entering the text at this time, we also used this unit. Between these Bunsetsu, there are some dependency relations called Kakari-Uke which can be decided uniquely for any sentence. We can consider that in case there is a Kakari-Uke relation between word A and B, A is modifying B.

This time we defined the distance between words mainly based on this Kakari-Uke relation, and then classified them into number of groups using some clustering techniques. As the result we got a base-dictionary which can represent Kakari-Uke relation between these groups.

It is expected that syntax, semantics or knowledge of the world can be naturally embeded in this dictionary and this type of lexicon is highly useful in the Japanese sentence analysis.

(2) Overview of ESSAY

ESSAY (Experimental System of Sentence AnalYsis) parses Japanese sentence by analyzing Kakari-Uke relation between Bunsetsu in input sentences.

This system has dictionary driven feature, and does not depend on usual syntactic and semantic models. Thus this system can be used for evaluation of dictionary, which is described in (1).

The input to this system is a Japanese sentence, which is segmented in Bunsetsu unit, and the output from this system is labelled binary tree structure, which represents syntactic structure of the input sentence.

The algorithm to extract this structure is very simple, and special linguistic knowledge is not embedded in the procedurable way. The decision of tree structure is based on Graph theoretic processing, and labelling of Kakari-Uke relation is processed by using Statistical decision theory.

As stated above, this system has its linguistic knowledge in the declarative way by the form of dictionary, thus structure of system is simple, and rich in modularity. But procedurable knowledge can be easily implemented, if we need it.

By taking this approach, it is possible to get the way to construct a flexible system, which has rich ability of adaptation to specified world. This point is one of the merits of our approach, in comparison with usual approaches, that tend to depend on researcher's framework.

In this paper, we present several experimental results which show the validity of our approach.