Implementation of an Automatic Paraphrase System

Osamu Sato

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

February 15, 1999

Keywords: paraphrase, rewriting, morphological analysis.

Various documents are made in digital form as the Internet environment becomes enough for practical use in recent years, we can get degital documents using World Wide Web(WWW) via Internet. However, many documents are not comprehensible for a lot of people. There are many documents which are written with domain specific expressions (e.g., technical terms and peculiar expressions), or are written in formal and rhetorical words.

In this research, I studies the method which paraphrases such incomprehensible documents into comprehensible, and implemented the system which paraphrases documents automatically.

What kind of paraphrase we need to do strongly depends on the object and the field of the targe document; I chose the report documents which are published from the government, the municipalities, and the official organizations because they contain a lot of difficult, formal and rhetorical —, expressions, and studied the paraphrase to make them easy enough to read.

I selected two documents: "Circulation of information rule (report) on the Internet (10764 byte, about 5000 words)" by Ministry of Posts and Telecommunications and "Ideal way of the new financial administration (7335 byte, about 3600 words) " by Ministry of Finance from WWW as targets, and paraphrased difficult expressions in them.

I paraphrased 162 examples, then examined them. Based on this examination, I designed the paraphrase system as follows:

1. Analyze a sentence morphologically, and make the result (the sequence of morphological units) the target for the paraphrasing.

2. What paraphrase to be done is described by a paraphrase rule, which replaces a subsequence of morphological units into another subsequence.

3. The system provides several subroutines, which enable us to write paraphrase rules easily.

The paraphrase system consists of the following six modules;

Preprocessing analyzes the input sentence morphologicallly, and outputs the results (the sequence of morphological units).

Paraphrase rule set is a program of the system. A rule indicates what paraphrase should be done.

Rule compiler converts the paraphrase rule set into the execution form.

Rewriting engine applies the paraphrase rule set in the execution form to the sequence of morphological units.

Subroutines are functions that can be used in paraphrase rules: e.g, inflection adjustment.

Dictionaries are used by subroutines.

The system works as follows:

1. Analyze the input sentence morphologically, and converts it into the sequence of morphological units. This is done by the preprocessing module.

2. Apply the paraphrase rule set to the morphological units. The paraphrase rule set is converted into the execution form previously by the rule compiler, and the rewriting engine does an actual rewriting by using the rule set in execution form.

3. Convert this output sequence of morphological units into the character string which is the final output.

I performed an experiment with this system. I used "Hashimoto Cabinet premier administration policy speech in the 136th pilgrimage association (28319 bytes about 14000 words)" as the target document. The system did 87 paraphrases; 63 paraphrases (72 percent) are correct, and 24 paraphrases (28 percent) are incorrect. Incorrect paraphrases include grammatical errors or the change of the meaning.

From this experiment, I obtained the following facts:

1. This system can fo the word-sequence to word-sequence replacement.

2. This system outputs the incorrect paraphrase sentence if the structural information of the sentence is required.

3. Even the output sentence is correct grammatically, it can be incorrect semantically.