EFFECTIVENESS OF COMBINING LEARNING RULES AND ANALOGY IN SEMANTIC ANALYSIS FOR JAPANESE UNKNOWN SENTENCES

Hideyuki SHIBUKI
High-Tech Research Center
Hokkai-Gakuen University
Minami 26, Nishi 11-1-1, Chuo-ku
Sapporo-shi, 064-0926 Japan
email: shibu@eli.hokkai-s-u.ac.jp

Yoshio MOMOUCHI
Graduate School of Engineering
Hokkai-Gakuen University
Minami 26, Nishi 11-1-1, Chuo-ku
Sapporo-shi, 064-0926 Japan
email: momouchi@eli.hokkai-s-u.ac.jp

Kenji ARAKI
Graduate School of Engineering
Hokkaido University
Kita 13, Nishi 8, Kita-ku
Sapporo-shi, 060-8628 Japan
email: araki@media.eng.hokudai.ac.jp

Koji TOCHINAI
Graduate School of Business Administration
Hokkai-Gakuen University
Asahi-machi 4-1-40, Toyohira-ku
Sapporo-shi, 062-8605 Japan
email: tochinai@econ.hokkai-s-u.ac.jp

ABSTRACT
In natural language analysis, it is one of the most difficult problems to analyze various sentences using limited knowledge. A Rule-based analyzer has the advantage of regular analysis using relatively fewer rules, and an Example-based analyzer has the advantage of flexible analysis using analogy. We propose a framework of hybrid analyzer that has both of the advantages. The hybrid analyzer learns rules from training data, and analogizes unknown sentences with the rules. We conducted experiments of assignment of semantic roles using three analyzers including the hybrid analyzer. In accuracy, the hybrid analyzer was the best result of them.

KEY WORDS
Natural Language Processing, Knowledge Acquisition, Temporal Reasoning, Case-based Reasoning

1. Introduction
In natural language analysis, it is one of the most difficult problems to analyze various sentences using limited knowledge. It is known that the following regularity and flexibility are observed in natural language. Regularity of language controls syntactic or semantic structure of a sentence systematically. Thereby, we can judge whether or not a sentence belongs to a language. However, in the language, there are many exceptional sentences not to be controlled regularly but to be understood flexibly. This ambivalent situation is a bottleneck in natural language analysis, and the following approaches have been done in order to solve it.

Rule-based analyzer, which is classical, has the advantage of representing regularity as relatively fewer rules, while it has the disadvantage of a lack of flexibility, that is, it is unable to analyze an unknown sentence that is not covered by the rules. Since it is hard to manually give a set of rules covering any sentences, researches for automatic acquisition of rules from sentences in a corpus to analyze the sentences have been done[1]. However, it is not realized so far to learn a set of rules covering any sentences, and facilities are needed to analyze unknown sentences flexibly. Statistics-based analyzer and Example-based analyzer have been studied for flexible analysis. Statistics-based analyzer[2] represents regularity as conditional probability, and can infer a probability of unknown sentence using smoothing technique. The disadvantages are that the method requires a large amount of sentences to calculate reliable probability and loss of most of linguistic information. Example-based approach has been studied popularly in machine translation, and there are a few researches in analysis[3]. An Example-based analyzer analyzes a sentence by analogizing the sentence to sentences analyzed in a corpus, and has the following advantages. It can flexibly analyze an unknown sentence using information of similar sentences in a corpus, and it can use linguistic information unlike a Statistics-based analyzer because sentences in a corpus are kept as they are. However, the disadvantage is that the method requires a large amount of sentences like a Statistics-based one. Since each approach has advantages and disadvantages, a hybrid analyzer that one advantage compensates for another disadvantage should be developed. We consider our hybrid analyzer as one combining Rule-based approach that can represent regularity by fewer rules with Example-based approach that can analogize a sentence flexibly. In other words, the hybrid analyzer keeps rules instead of examples, and analogizes unknown sentences with the rules. We have studied the way of learning rules for effective analysis. We consider that learning is necessary from the following reasons. Language has always changed, and unimproved knowledge will not cope with new words or expressions. The improvement should be performed automatically by learning because manual improvement requires vast labor. Therefore, knowledge our analyzer uses is automatically acquired from sentences by minimally supervised learning. Note that acquisition
and analogy are performed using the same sentences. In section 5, we demonstrate that the performance of hybrid method is better than the performance by Example-based approach using the same sentences.

Our analyzer is developed for semantic analysis that is to assign semantic role between two phrases such as 'AGENT' or 'MANNER'. Knowledge for semantic analysis is not firmly established relatively to morphological or syntactic analysis, and manual establishment of the knowledge tends to be subjective. It is worthwhile learning the knowledge automatically by objective criterion. We would like to acquire the knowledge from raw sentences finally. However, it is so hard at present that we use sentences with annotation of dependency between phrases from the following reasons. There are  

1. In this paper, we defined a phrase as a unit that consists of a lexical word and zero or more than grammatical words. For example of a sentence "The boy goes to school.", phrases are "the boy", "goes" and "to school". The detail is described in section 2.

2. The annotation has no concern with parts of speech. The annotation for a word includes only information whether the word is lexical or grammatical.

3. The EDR corpus is one of the most popular corpora used in Japanese natural language processing. It includes about 210,000 sentences given syntactic and semantic annotation.

first row expresses a sentence written in Japanese, the second is the pronunciation, and the third and the forth are the meanings of each word\(^5\) and the meaning of the sentence in English respectively. For convenient explanation, each word is divided with space, lexical words are highlighted, and a colon expresses a border between *bunsetsu*’s. Since some grammatical words\(^6\) work significantly in order to determine a semantic role, a *bunsetsu* order is free relatively to a word order in English except a restriction that a *bunsetsu* depends on a backward *bunsetsu* in general. For example, a sentence “He looks at her.” is expressed by two sentences in Japanese shown as (a1) and (a2) in Figure 1. Japanese native speakers can resolve both of the sentences into the same dependencies (a3) and (a4), and can understand each dependency independently. Therefore, our analyzer determines a semantic role of a dependency independently.

The effect of grammatical words is significant but not determinate. There are problems in the following cases: that different grammatical words have the same semantic role like (b1) and (b2), or that the same grammatical word has different semantic roles like (c1), (c2), and (c3). To solve them, we consider that the former needs clustering of grammatical words translated to the same semantic role, and that the latter needs analysis using lexical words in a dependency. Therefore, we use clusters of grammatical words and lexical words for determination of semantic roles.

### 3. Semantic Analysis

It is difficult to express a result of semantic analysis using only sentences with syntactic annotation because semantic roles never appear in the sentences. We introduce an intermediate symbol between an input sentence and a semantic role, and divide the task of semantic analysis into two subtasks shown in Figure 2. The first task is to classify an input sentence into an intermediate symbol\(^7\), and the second is to translate the symbol into a semantic role. If a symbol corresponds to a semantic role one-to-one, performance of analysis depends on classification into an appropriate symbol in the first task. By separating performance of semantic analysis from the expression of output, we actualize semantic analysis using syntactic annotation. We call the first task as semantic analysis in this section.

#### 3.1 Rule-based analyzer

Conditions of applying a rule for determination of a semantic role are grammatical and lexical words described in section 2. For example of a rule that [animate]\(^5\) can be

\[ \text{‘AGENT’ of [act], a sentence “She goes,” is analyzed as follows. Because “she” and “go” are kinds of [animate] and [act] respectively, the rule can be applied to the sentence. Therefore, the analyzer can determine that “she” is ‘AGENT’ of “go”. Even if an intermediate symbol is used instead of a semantic role, the conditions are never changed. Thesaurus is generally used to judge whether or not a word belongs to a concept.} \]

#### 3.2 Example-based analyzer

A simple Example-based semantic analyzer is designed as follows. For example of two similar sentences “He goes.” and “She goes.”, if a semantic role of “he” to “go” is ‘AGENT’, the analyzer can infer that a semantic role of “she” is also ‘AGENT’. Even if the analyzer is not given what intermediate symbol corresponds to a semantic role, it can judge that “he” and “she” have the same symbol. Therefore, it outputs a symbol of an input sentence as a symbol of the most similar sentence to the sentence.

Since our analyzer uses grammatical and lexical words, a similarity between dependencies is calculated based on similarities between the words in the dependencies. Thesaurus is generally used to calculate a similarity between words, and we calculate a similarity \(sim_w\) between words, \(w_1\) and \(w_2\), by the expression (1).

\[
\text{sim}_w(w_1, w_2) = \frac{2 \times \text{depth}(cw)}{\text{depth}(w_1) + \text{depth}(w_2)}, \quad (1)
\]

where the \(\text{depth}(w_1)\) is the number of nodes from the root to \(w_1\) in a thesaurus, the \(cw\) is the deepest common hyponym of \(w_1\) and \(w_2\). Since we would like to normalize a similarity to avoid the influence of size of thesaurus, we use the expression that can represent a similarity in the range from 0 to 1. The research of better expression is a future work. A similarity between dependencies is a sum of similarities between words in the dependencies multiplied by coefficients. In this paper, the coefficient for grammatical words is 5, and one for lexical words is 1, because grammatical words are more significant. The research of better coefficient is a future work.

Although a general Example-based analyzer uses a threshold of similarity for accuracy, our Example-based analyzer does not use the threshold because our purpose in this paper is to analyze unknown sentences at any cost.

#### 3.3 Hybrid analyzer

At first, we sum up the knowledge that the Rule-based and the Example-based analyzers need. The Rule-based analyzer needs rules and thesaurus, and the Example-based analyzer needs sentences and thesaurus. If the rules can be acquired from the sentences, both analyzers need the same knowledge. Therefore, our hybrid analyzer also needs the same knowledge by implementing a

\[ \text{In this example, an English sentence that consists of only two phrases is used for convenience. There is no problem in cases that a Japanese sentence consists of more than two *bunsetsu's, because the sentence can be resolved into dependencies from the reason described in section 2. Thus, we use a sentence as a meaning of dependencies.} \]

\[ \text{A word in parentheses means grammatical meaning.} \]

\[ \text{In actual, a *bunsetsu* may include several grammatical words, and a grammatical word in the partnered *bunsetsu* in a dependency may influence a semantic role like the passive voice. Thus, our analyzer uses a group that includes all grammatical words in a dependency. For convenience, we call the group a grammatical word simply.} \]

\[ \text{The symbols are acquired by the way described in section 4.} \]

\[ \text{A word in brackets means an abstract concept.} \]
method for automatic acquisition of rules from sentences described in section 4.

Figure 3 shows the difference between our hybrid analyzer and the Example-based analyzer. The hybrid analyzer searches rules that can be applied to an input sentence at first like the Rule-based analyzer. If no rule can be applied, the analyzer analogizes the sentence to not examples but rules. Analogy to a rule is the difference from the Example-based analyzer. This is performed by regarding conditions of the rule as a sentence. For example of a rule that [animate] can be 'AGENT' of [act], it is regarded as "[animate] [act]."\(^{10}\) Because both a word in the input and a concept in the rule are in the same thesaurus, their similarity can be calculated by the same way described in 3.2. If a word in the input is unknown, the analyzer calculates the similarity by regarding the word as the root of the thesaurus. This is because of avoiding to bias against the word, and because the root is the neutral node from all words.

4. Learning

We describe only the outline of learning because of the limitation of space of the paper. The first task in Figure 2 needs intermediate symbols, thesaurus, and rules for determination. The second task needs rules for translation. The knowledge used in the first task is acquired from sentences with syntactic annotation, and the rules for translation are from sentences with semantic annotation.

We assume creation of intermediate symbols to be clustering of dependencies that would have the same semantic role. Therefore, intermediate symbols are acquired as follows. At first, a unique initial symbol is assigned to each dependency in training data. Next, a hierarchy of initial symbols is constructed using the way to construct a thesaurus described in the next paragraph. Then, our system descends the hierarchy from the root to the daughter nodes until the number of the nodes is equal
to the number of semantic roles. Finally, the nodes are set as intermediate symbols.

Three types of thesauri are constructed for grammatical words\(^{11}\), lexical words, and intermediate symbols. A type of thesaurus is constructed using the other types of initial thesauri of which the root includes words or symbols directly. This is based on an assumption that similar symbols or words tend to occur with other similar words or symbols. For example of grammatical words, the thesaurus is constructed based on an assumption that similar grammatical words tend to occur with similar lexical words and symbols. Similarity between lexical words or symbols is a value calculated by expression (1), and similarity between grammatical words in two dependencies is a sum of a value between lexical words and a value between symbols in the dependencies. The thesaurus is created by bottom-up constructing a pair of grammatical words in order of the high similarity.

Rules are acquired by generalizing words in examples using the above-mentioned thesaurus. If two examples are classified into the same intermediate symbol, a rule can be acquired by replacing words in the examples with a common hypernym of the words shown in Figure 4. An acquired rule can be generalized with other examples or rules again. However, training data may include inappropriate examples for acquisition of rules like broken expressions. Such inappropriate examples need to be removed. Furthermore, it is natural that there are several rules to classify the same symbol without generalization such as a rule for an object in the active voice and a rule for a subject in the passive voice. Such rules need to be kept as they are. Therefore, we considered three ways of learning rules that are generalizing, removing, and keeping examples and rules. Since the ways influence the performance of analysis, rules are learned using the best way to analyze the training data.

An intermediate symbol is mapped to a semantic role as follows. At first, our analyzer analyzes sentences annotated with semantic roles. The result is used to count

\(^{10}\)Because a rule is learned by generalizing words in sentences with keeping the position, a rule can be translated to a sentence.

\(^{11}\)This is used to judge grammatical words translated to the same semantic role from the reason described in section 2..
6. Consideration

Table 1 shows the result. The result shows that the hybrid analyzer used the training data more effectively than the other analyzers. It is interesting that the different results are derived from the same training data.

The results of the hybrid and the compound analyzers are better than the result of the Example-based analyzer. The reasons are that there are not enough examples for the effective Example-based analysis, and that analysis using rules is determinate unlike analogy using examples. When a concept in a rule is generalized from words in examples, the concept can include the other words. If an included word is correct, the hybrid and the compound analyzers can analyze sentences with the word correctly and determinately. However, the most similar sentence inferred by the Example-based analyzer is not always correct because of influence of other words in the input sentence.

We consider the reason why the hybrid analyzer is superior to the compound analyzer. Because the difference between the analyzers is what is used for analogy, the result shows that analogy with rules is more effective than analogy with examples in the case of analyzing unknown sentences. In the case of unknown word, the analyzers analogize the root of a thesaurus to a concept in a rule. The more words the concept includes, the more similar the concept and the root are in the thesaurus. Therefore, the hybrid analyzer analyzes unknown sentences using a rule with concepts including more words, and the result was more correct than the result using examples. This may be explained by the similarity-coverage model[6]. The model explains inference of human beings. As a factor to judge whether or not an attribute of an object is true in the case that the attribute of other objects is true, the model gives coverage of a category that is how many objects the category includes. Human beings tend to judge that inference is correct if the coverage is high. This is similar to the mechanism that the hybrid analysis selects a rule with concepts including more words. The effectiveness of the hybrid analysis may originate from the similar mecha-
Table 1. Result of the experiment.

<table>
<thead>
<tr>
<th>Analyzer</th>
<th>No. of correct dependencies</th>
</tr>
</thead>
<tbody>
<tr>
<td>The Example-based analyzer</td>
<td>633</td>
</tr>
<tr>
<td>The compound analyzer</td>
<td>658</td>
</tr>
<tr>
<td>The hybrid analyzer</td>
<td>718</td>
</tr>
</tbody>
</table>

nism. However, the research of the relationship to inference of human beings is a future work.

Hence, we conclude that the hybrid analyzer is effective in the situation of analyzing unknown sentences using limited sentences.

There are the following previous researches related to hybrid methods that are not to determinate semantic roles. The research [4] acquired rules to increase coverage of examples in Example-based machine translation. Their method is similar to our Rule-based method without analogy, and the coverage is increased to about 90% using 1 million of words of training data. However, the result also shows that the method cannot process all sentences unlike our hybrid analyzer. The research [5] proposed a method to direct phrasal attachments using large-scale lexical knowledge. Therefore, the purpose is different from ours that is effective use of limited linguistic resource.

7. Conclusion

We have proposed a hybrid semantic analysis combining Rule-based analysis with Example-based analysis. The hybrid analysis has the advantages of regular analysis using relatively fewer rules and of flexible analysis using analogy. Therefore, the hybrid analyzer is effective in the situation of analyzing unknown sentences using limited sentences. We would like to apply the framework of hybrid analysis to other natural language processing systems.

Acknowledgement

This work is partially supported by the High-Tech Research Center of Hokkai-Gakuen University.

References


