Better Translation with Knowledge Extracted from Source Text

Satoshi KINOSHITA, Miwako SHIMAZU, Hideki HIRAKAWA
R & D Center, Toshiba Corporation
1 Komukai-Toshiba-cho, Saiwai-ku, Kawasaki, 210 Japan
E-mail: kino@isl.rdc.toshiba.co.jp

Abstract

This paper presents a framework in which a source text is translated using not only given knowledge but knowledge which is extracted from the text. First, co-occurrence relations between words are extracted prior to translation, and then are used in translation to resolve ambiguities in the source text. The important feature of the system is that extracted relations are categorized as either "reliable" or "semi-reliable" according to ambiguities in the analysis. "Reliable" relations are highly likely to be correct and therefore can be used for translating other texts. By contrast, "semi-reliable" ones may contain some errors and therefore should be restricted to texts from which the relations have been extracted. However, "semi-reliable" relations prove valuable when the system has limited knowledge; in our experiment the accuracy of disambiguating verb-noun dependencies has improved from 91.5% to 93.0% with a use of both "reliable" and "semi-reliable" knowledge, while with only "reliable" knowledge the accuracy recorded 92.3%. The effectiveness was furthermore examined when the system was provided with manually collected co-occurrence relations which are essentially equivalent to a complete set of co-occurrence relations in the domain.

1 Introduction

Semantics is useful for resolving ambiguities in natural language. For example, selectional restrictions on verbs, usually represented in the form of semantic markers or concepts in a sophisticated thesaurus, give positive/negative preference to possible interpretations. Another example is to recognize correct coordinate structures by calculating semantic similarities between words. With complete semantic information, syntactic noises, which often impedes operational machine translation(MT) systems, can be overcome.

However, analysis in the current MT systems is not centered on semantics because practically MT developers cannot afford to provide users with complete semantic information. Consequently syntax, which is relatively easier to handle, takes over semantics. Suffice it to say that syntactic variations between domains are relatively small, while there are significant differences in the senses of words; a different set of semantic information should be prepared for each specific domain because every subject domain has its characteristic semantics.

To incorporate domain-specific semantic knowledge into the system, research in [1][2]

[5][6] has proposed to develop a program which extracts semantic information from corpora. But this approach poses the following problems, which make it unrealistic in practical environments. First, a large amount of text in a specified domain is required, which is not always available, particularly at a single user site. Second, texts to be fed to an extraction program often need to be tagged or annotated. This tagging is a time- and resource-consuming task and is not easily accepted by potential MT users. To reduce the burden of tagging, a modified approach has been proposed, in which man and a computer cooperatively produce semantic knowledge from an untagged corpus[7].

These approaches all aim at producing complete semantic knowledge for the specified domain; huge corpora and human involvement are necessary for building up semantic information with broad coverage and high accuracy.

Another approach for utilizing domain-specific semantic knowledge has been proposed, in which no corpus is required. In this approach, the system uses semantic information extracted from texts which it analyzes or translates[4][8]. Such semantic information would never be equivalent to what could be extracted from huge corpora. But this approach seems plausible because knowledge for resolving ambiguities in a text is likely to be found in the same text.

In this paper, we propose a framework in which translation is done based on this approach; semantic information is extracted from a source text before translation and is later used for translating the very text. As the extracted information accumulates, semantic information will gradually increase. This approach is motivated by our guiding principle that reduction of human involvement in the extraction process, like preparing sample texts and checking extracted data, is far more important than the accuracy or completeness of extracted information; even if extracted information contains some errors, we think it is reasonable as long as it serves the purpose, namely improvement in translation.

In the following sections, we will first describe the framework in detail, and then describe the experiment results which show its effectiveness.

2 Framework

2.1 Two-phase Translation

Figure 1 shows an overview of the framework. The translation process consists of two phases. In the first phase, the source text is analyzed using the analysis module of the translation system, and semantic information is extracted from the analysis result of each sentence and is added to the knowledge base. In the second phase, translation is carried out in three sub-phases: analysis, transfer and generation. In the analysis, information extracted in the first phase is used together with the given syntactic/semantic knowledge for resolving ambiguities.

Note that two ways are possible with respect to the timing of extraction and use of extracted information. One is to carry out machine translation and extract information

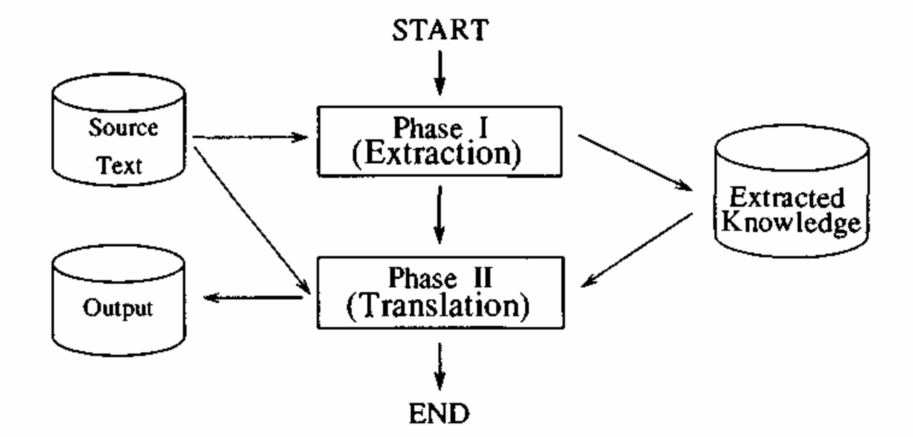

Figure 1 Overview of Translation Process

simultaneously; the other is to translate a text once the extraction has been done for the entire text. We have adopted the latter, since, in the former, information available for parsing would be limited to the preceding context, resulting in the poverty of information.

One potential disadvantage of this approach is that sufficient amount of knowledge for resolving ambiguities might not be extracted from a source text because only a small amount of "reliable" semantic information is extracted from a single sentence. Here "reliable" means that information is extracted from an analysis result, or part of it, which is determined uniquely or unambiguously. For example, if a word modified by a prepositional phrase is uniquely determined as in the sentence (1), a modification relation between the modifier and the modifiee, "see with a telescope" in this example, can be extracted as a "reliable" relation.

(1) I saw it with a telescope.

Most extraction programs proposed so far are designed to extract only "reliable" information from given texts because human checking on the extracted data is not presupposed or is to be kept as small as possible. That is why a huge corpus is necessary to extract a sufficient amount of semantic information for a specific domain. However, few user can prepare such a huge corpus for this purpose.

To overcome this problem, our framework utilizes "semi-reliable" information which is less dependable than "reliable" one. As described below, "reliable" information is accumulated for the domain knowledge, whereas "semi-re liable" one is used only for translating texts from which it is extracted.

The above framework has been applied to Japanese-English machine translation. In the subsequent sections, attention is given to Japanese linguistic phenomena.

2.2 Extraction of Co-occurrence Relations

Semantic information which we focus on in this paper is co-occurrence relations to be categorized as below:

i) co-occurrence relations between a verb and its case filler
ii) co-occurrence relations between two nouns

Co-occurrence relations are extracted from a semantic tree, namely an output of the sentence analysis, by traversing every node in the tree and checking its modifiers.

i) Co-occurrence relations between a verb and its case filler

This type of co-occurrence relation is extracted by searching a modifying relation between a postpositional phrase and a verb. To illustrate, two co-occurrence relations, i.e. (3a) and (3b), will be extracted from the sentence (2).

(2) sono ato, teepu-wo heddo-ni makitsukeru.
that after tape-OBJ head-around wind
'After that, Φ winds the tape around the head.'

(3a) teepu-wo makitsukeru ('wind a tape')
(3b) heddo-ni makitsukeru ('wind around a head')

Extracted relations are categorized as either "reliable" or "semi-reliable" according to the analysis results. When a relation between a noun and a verb is determined unambiguously, it is categorized as "reliable." Ambiguity of a modifying relation between a postpositional phrase and a verb is checked with respect to two points:

-dependency: If a postpositional phrase syntactically modifies more than one verb, the "dependency" relation is ambiguous.
-semantic roles: If a postpositional phrase modifying a verb can be a candidate of more than one case filler of the verb, the "semantic" relation is ambiguous.

If the relation does not meet these requirements, it is regarded as unambiguous.

In the above example, the co-occurrence relation (3a) is categorized "reliable" because a verb that the postpositional phrase "teepu-wo" modifies and the semantic relation between them are determined unambiguously.

On the other hand, when there is an ambiguity, the system's final interpretation may be wrong, so is the co-occurrence relation based on the result. Accordingly the extracted relation is categorized as "semi-reliable."Consider the sentence (4).

(4) kono taimingu-de sutaatopojishon-ni idou-saseru.
this timing-on starting position-to move-cause
'On this timing, Φ moves Φ to the starting position.'

In this sentence, the verb that the phrase "sutaatopojishon-ni" modifies is determined unambiguously because "idou-saseru" is the only verb after it. However, there exists an ambiguity for a semantic interpretation of the postposition "ni" because the functional word "saseru" adds a semantic role to this postposition; the noun "sutaatopojishon" can be an object which moves or a destination of the "moving" event. In such a case, heuristic knowledge is applied to determine whether the relation is plausible.

Extracted relations categorized as "semi-reliable" may contain some errors, while all "reliable" relations are highly likely to be correct. This difference in reliability leads to different types of usage of these relations; "reliable" relations can be used for translating other texts, whereas "semi-reliable" relations should be restricted to texts from which the relations are extracted.

ii) Co-occurrence relations between two nouns

For this type, we are concerned with modifying relations where the postposition of a modifying postpositional phrase is "no"('of'). When the word that such a postpositional phrase modifies is determined unambiguously, the relation between two nouns will be extracted. In the sentence (5), the postpositional phrase "teepu-no" can modify only the noun "tenshon," and the co-occurrence relation (6) is extracted.

(5) teepu-no tenshon-niyori, maikurosuicchi-ga on-ni naru.
tape-GEN tension-due to microswitch-SUBJ on-to become
'Due to the tension of the tape, the microswitch will be on.'

(6) teepu-no tenshon ('tension of a tape')

Every co-occurrence relation between two nouns is considered "reliable" because it is extracted only when a syntactic dependency between them is uniquely determined and no semantic role of the postposition "no" is taken into account.

2.3 Application of Extracted Knowledge

Figure 2 presents the sentence analysis flow. An input word sequence (output of morphological analysis process) is analyzed by the three processes: syntactic analysis, semantic-dependency generation and optimal tree search. The output of each process is called a syntactic-dependency tree(SYD-tree), a semantic-dependency graph(SED-graph) and a semantic tree(SEM-tree), respectively. Information extracted from source texts is referred to in the second process.

(1) Syntactic analysis

The syntactic analysis process generates only one SYD-tree from an input word sequence using an extended CFG parser. Figure 3 shows the SYD-tree for the Japanese sentence (7).

(7) kairo-wa teepu-wo monitaa-suru
circuit-TOP tape-OBJ monitor
’The circuit monitors the tape.'

In a SYD-tree, a node corresponds to a word in the input sentence and an arc indicates a syntactic dependency relation (case, nominal-modification etc.). All candidates for

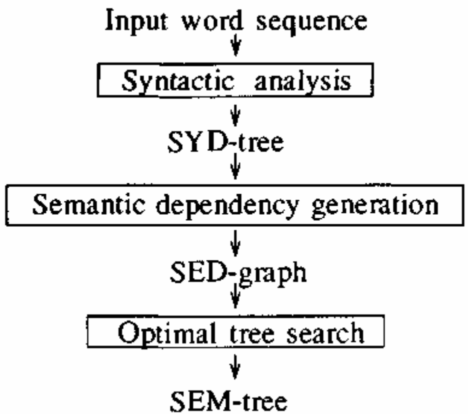

Figure 2 Flow of Sentence Analysis

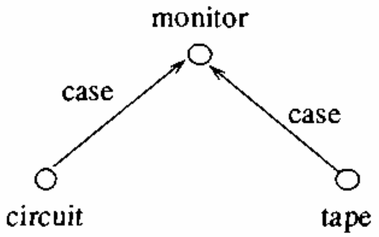

Figure 3 Syntactic Dependency Tree

one node's dominator are in its ancestor nodes. In this sense, a SYD-tree implicitly holds all possible syntactic dependency structures.

(2) Semantic-dependency generation

The second process generates a SED-graph from a SYD-tree. The SED-graph expresses possible dependency and semantic relations between words in a sentence, A directed arc indicates one possible semantic dependency relation between two words. Arcs are given labels to indicate the semantic role of modifying words. SED-arcs are set by picking up two nodes in a SYD-tree along with a path in the tree and assigning possible semantic relations between the two nodes (SED generation rule). For example, the SED generation rule for the postposition "wa" generates two SED-arcs ("agent" and "object") between "kairo"('circuit') and "monitaa-suru"('monitor') because the postposition "wa" can mark these two different semantic roles(See Figure 4).

Each arc has a priority score called a weight which is given by referring to syntactic and semantic knowledge. For example, if a semantic marker of a case filler satisfies the selectional restriction of a verb, *a* positive score is added to a SED-arc which represents the semantic relation while a negative score is added if the restriction is violated.

Co-occurrence relations extracted from the source text are used as well as semantic knowledge represented with semantic markers. If a semantic relation which is represented by a SED-arc is stored as an extracted relation, the arc is given a score, which is higher than a score given in a semantic test with semantic markers. When scores based on semantic markers and extracted relations contradict each other, results using the latter knowledge are used.

In the current implementation, there is no difference between "reliable" and "semi-reliable" information regarding a score which is given by referring to co-occurrence relations. Also, frequency of extracted relations is not taken into account. More elaborate scoring is left for the future work.

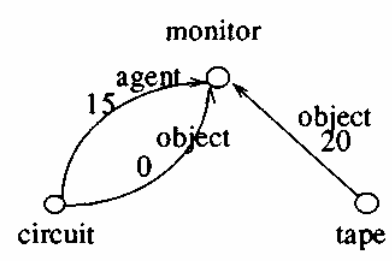

Figure 4 Semantic Dependency Graph

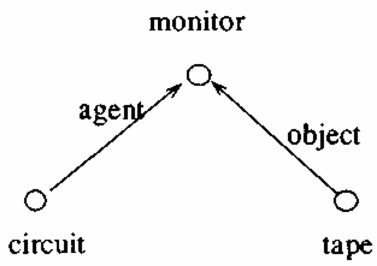

Figure 5 Optimal Semantic Tree

(3) Optimal tree search

A semantic tree(SEM-tree) of a sentence which is extracted from a SED-graph is defined as a well-formed spanning tree. Semantic trees must satisfy the following constraints:

Constraint #1 Cross-Dependency
No two arcs cross in a tree.
Constraint #2 Multiple Valence-Occupation
No two arcs in a tree occupy the same valence of a predicate.

The weight of a SEM-tree is the sum-total of weights of arcs in the tree. A SEM-tree which has the maximum weight is called the optimal tree, which will be the final output of the sentence analysis(See Figure 5).

3 Experiment

3.1 Methodology of the Experiment

Because extracted information mentioned in the previous chapter is used for resolving ambiguities in the analysis phase, the effectiveness of the framework should be evaluated measuring the rate of correct parsing. Although any change in the parsing precision is directly reflected in translations, comparison of translation results will not give a true and full account of parsing. In the following experiments, we measured the parsing accuracy based on the verb-noun modification relationship; if a noun, more precisely a postpositional phrase, syntactically modifies more than one verb, the relation between the noun and the verbs is extracted and is checked whether the modification is correct or not.

Our experiments were done on a Japanese manual of an electric appliance which contains 1,290 sentences. To begin with, the text was analyzed using the analysis module of our MT system ASTRANSAC[3] with its standard dictionary and grammar to extract semantic information explained in the previous chapter. In parallel, ambiguous verb-noun modification relations were extracted. Next, the text was analyzed again to extract these modification relations, this time using the extracted information. The extracted relations were compared to evaluate how much improvement was gained in the precision of sentence parsing.

3.2 Experimental Results

3.2.1 Extracted Data

Table 1 shows the number of extracted items classified by their types.

The first and second rows, namely (1) and (2), indicate the number of extracted co-occurrence relations mentioned in the previous chapter. The left columns give the total number of items, and the right columns, the number of unique occurrences. Extracted relations in (1) are divided into two types; reliable and semi-reliable relations.

The third row (3) indicates the number of words which have ambiguous modifying relations; there are more than one directed arcs extending from a node that corresponds to the word.

On the basis of extracted data categorized as (3), we analyzed the parsing precision of the current MT system, focusing on the modifying relations. Table 2 presents the results.[1] Of 292 errors of parsing modifying relations, verb-noun relations amount to 84, which account for 29% of the total errors. They top the list for possible causes of parsing errors. Among words which have ambiguous modifying relations, 991 belong to the verb-noun relation group. So the accuracy of analyzing the ambiguous verb-noun relations is 91.5%.

3.2.2 Effectiveness of the Data

We examined the change in the rate of correct sentence parsing by applying the extracted information to the MT system in two different experiments. Specifically, we observed to what degree the precision of parsing verb-noun modifying relations improves when we used only "reliable" semantic data and when we used both "reliable" semantic data and "semi-reliable" semantic data.

In the first case, 16% of incorrect parses were corrected, while some of the originally correct ones were incorrectly parsed, which in number is an equivalent of 7% of the incorrect parses. In contrast, when "semi-reliable" data was additionally used, 25% of the incorrect parses improved, but the same percentage of the parses as in the first case worsened. This means that, of the total of 84 errors involving verb-noun relations, 9% of the errors have been corrected with the use of "reliable" semantic data and 18% improved with the additional use of "semi-reliable" data.

We then calculated the overall accuracy rate of the system. Thus, we would get the following rates according to the types of semantic information used:

[1] Although the number of words which have ambiguous modifying relations totaled 2,354, we omitted 34 of them as inappropriate. Among them are input errors of the original text, marks which are outside the scope of the linguistic analysis, and errors of determining word boundaries or part of speech due to incorrect morphological analysis.

Table 1 Statistics of Extracted Items from Test Text

	number in total	number of unique occurrences
(1) verb-noun(reliable)	1,567	838
verb-noun(semi-reliable)	432	320
(2) noun-noun	483	358
(3) ambiguous modification	2,354	

Table 2 Result of Analyzing Modifying Relations

Type of modification	correct	wrong	total
verb-noun	907	84	991
others	1121	208	1329
total	2028	292	2320

-No extracted data is used	91.5%
-Only reliable semantic data is used	92.3%
-Both reliable and semi-reliable data are used	93.0%

Thus, by the use of the extracted semantic data, the precision of parsing ambiguous verb-noun relations improves steadily, relative to the amount of information available.

(8) is a typical example of those which were correctly parsed with the use of extracted semantic information.

(8) den'atsu-ga sosa-chu-ni 1-kara 0-ni ochi-ta.
voltage-SUBJ operation during 1 from 0 to dropped
'Voltage dropped from 0 to 1 during operation.'

In (8), when no extracted information is used, the noun "den'atsu" is analyzed as the subject of the verb "sosa" because of their contiguity, obtaining the wrong translation "It dropped from 0 to 1 during voltage operation." However, since the co-occurrence relation "den'atsu-ga ochiru"('voltage drops') is included in the extracted information, the noun "den'atsu" is preferably analyzed as the subject of the verb "ochiru," thereby obtaining the correct translation as shown above.

On the other hand, (9) gives an example of erroneous parsing.

(9) buhin-wo toritsuke bisu-de kotei-suru.
parts-OBJ mount screws-INST fix
'Mount the parts and fix them with screws.'

In the above, syntactically two interpretations are possible with respect to the noun "buhin."

In the correct interpretation, the noun "buhin" would be shared by the two verbs "toritsuke-ru" and "kotei-suru." When no extracted information is used, the second verb correctly takes the noun "buhin" through the application of the system's general rules on parsing. However, since the extracted information contained the semantic relation "buhin-wo toritsukeru" ('mount the parts') but not the semantic relation "buhin-wo kotei-suru"('fix the parts'), the use of the extracted information would offer preference to the first semantic relation and the noun "buhin" is no longer recognized as the object of the verb "kotei-suru."

4 Discussion

4.1 Comparison with Other Method

The experiment results described in Chapter 3 shows that 25% of the incorrect parses have improved by utilizing the extracted information. We examined all cases individually where no betterment was achieved and categorized them into three classes with respect to knowledge for improvement, as illustrated in Table 3.

The "syntactic" problems are supposed to be conquered by adding syntactic knowledge. If an incorrect parsing is corrected by adding a preference rule for recognizing coordinations based on only syntactic clues, this error is classified in this category. Note that some of the errors of this type might be decreased with appropriate semantic information, but we categorized them as this type if they are likely to be corrected with additional syntactic knowledge.

The errors which fall into the "semantic" category are those judged to have potentials for improvement just by providing verb-noun co-occurrence relations. Therefore, improvement is expected simply by supplementing the necessary semantic data.

The errors classified as "knowledge-based" are those we judged require information involving multiple cases, which cannot be represented in a single co-occurrence relation. Also, those which require reasoning and the recovery of omitted words from context, known as zero-anaphoras, are treated as "knowledge-based."

"Semantic" errors are supposed to be reduced if complete domain-specific semantic information is available, but it is unknown how much of the errors will be eliminated in the current framework. To answer this question, we conducted another experiment. First, we manually checked all possible combinations of all nouns which are ambiguous with respect to the verbs they modify and all verbs listed as candidates for words these nouns modify. Next, we registered those which we think have co-occurrence relations into the database and carried out the same parsing experiment as done earlier.

This semantic information could be termed "limiting semantic information" as it were. This is essentially equivalent to a knowledge database that could be acquired from a huge corpus in the specific domain. In other words, a result which will be obtained using this information can reveal the limitation of the framework.

As a result of making use of "limiting semantic information," 45% of the incorrect

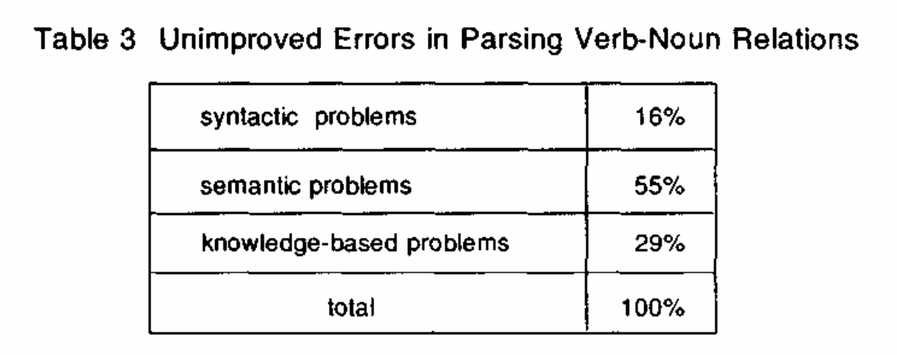

Table 3 Unimproved Errors in Parsing Verb-Noun Relations

syntactic problems	16%
semantic problems	55%
knowledge-based problems	29%
total	100%

parses improved, while some of the originally correct ones, which in number are equivalent to 11% of the parsing errors, worsened[2]. This in turn amounts to the accuracy rate of 94.5%. This shows that the precision of parsing verb-noun pairs can be supposedly upgraded to this upper limit utilizing semantic information of the form extracted in this experiment[3].

Furthermore, comparing the accuracy rate attained through the use of "limiting semantic information" with the value obtained earlier using semantic information extracted from the source text, 52% of the errors which have been corrected with the "limiting" information are covered with the use of extracted semantic information.

4.2 Extension

To use extracted co-occurrence relations effectively, generalization of the relations are essential. One apparent direction of generalization is to represent the relations at the level of semantic classes[4]. But as we mentioned earlier, such semantic classes are heavily dependent on subject domains, and therefore cannot be available beforehand. This entails constructing domain-specific semantic classes after translating a large number of documents in the same field and accumulating a reasonable amount of extraction data from them. For example, a set of nouns which can be objects of one particular verb may be said to belong to one semantic class. By introducing the semantic groups formed in this way, we can generalize our data efficiently. Needless to say, they will serve for recognizing correct coordinations.

[2] The worsened cases are categorized into two groups; 56% are due to sharing of objects mentioned in section 3.2.2 and 44% are due to sharing of subjects. We may argue that these are not so grave mistakes since they capture the correct semantic relations but were incorrect only in assigning a narrow scope. In the latter case, it is doubtful whether we should regard them as worsened examples, because they have subtle interpretations.

[3] Note that, if the system is provided with complete semantic knowledge, a preference score mentioned in 2.3 could be increased; this in turn makes influence of preference with syntax smaller, and could improve more incorrect parses.

5 Conclusion

We have attempted to resolve ambiguities which arise in the course of machine translation by using semantic information(co-occurrence relations) which was extracted from the source text to be translated, prior to translation. The motivation behind this is the present situation where it is difficult for MT users to utilize domain-specific semantic information. Our experiment on approximately 1,300 Japanese manual sentences resulted in the improvement of precision of parsing verb-noun relations from 91.5% to 93.0%. The present method has proved robust and efficient enough for practical use even without using thesauri, considering that the accuracy rate when "limiting semantic information" is used has reached 94.5%.

The important feature of our method is that extracted data is categorized into two types: 100% sure data("reliable" semantic information) and data which may contain some errors and therefore are less dependable("semi-reliable" semantic information). The reliable semantic information can be used in translating other documents as long as they belong to the same subject field. Thus, by accumulating this type of information, large-scale knowledge database in that particular field can be automatically constructed. Meanwhile, the application of semi-reliable semantic information should be restricted to the text from which the data is extracted. Our experiment confirmed that semi-reliable semantic information is helpful as heuristics to be used within the source text, although application of this type of data accompanies some risks in the sense that it is less reliable.

References

[1] Calzolari, N. and Bindi, R.: Acquisition of Lexical Information from a Large Italian Corpus, Proc. COLING 90, Vol.3 pp.54-59, 1990.

[2] Grishman, R. and Sterling, J.: Acquisition of Selectional Patterns, Proc. COLING 92, pp.658-664, 1992.

[3] Hirakawa, H., Nogami, H. and Amano, S.: EJ/JE Machine Translation System ASTRANSAC-Extensions toward Personalization, Proc. MT SUMMIT-III, pp.73-80, 1991.

[4] Inagaki, H., Miyahara, S., Nakagawa, T. and Obashi, F.: Sentence Disambiguation by Document Oriented Preference Sets, Proc. COLING 90, pp.183-187,1990.

[5] Nagao, K.: An Approach to Practical Semantic Processing, Proc. 39th Annual Convention IPS Japan, pp.652-653,1989 (in Japanese).

[6] Nakajima, H. and Kaji, H.: Automatic Acquisition of Knowledge of Word Co-occurrence from Text, Proc. 38th Annual Convention IPS Japan, pp.325-326, 1989 (in Japanese).

[7] Sekine, S., Ananiadou, S., Carroll, J. J. and Tsujii, J.: Linguistic Knowledge Generator, Proc. COLING-92, pp.560-566, 1992.

[8] Tanaka, K., Nogami, H., Hirakawa, H. and Amano, S.: Machine Translation System Using Information Retrieved from the Whole Document, Proc. 40th Annual Convention IPS Japan, pp.405-406, 1990 (in Japanese).