An Environment for Extracting Resolution Rules of Zero Pronouns from Corpora

Hiromi Nakaiwa
NTT Communication Science Laboratories
2-4 Hikaridai, Seika-cho, Souraku-gun, Kyoto 619-0237 Japan
nakaiwa@cslab.kecl.ntt.co.jp

Abstract

This paper proposes a practical integrated environment for extracting rules for the anaphora resolution of zero pronouns from monolingual and/or bilingual corpora. This method takes into account the practical situation for making resolution rules of zero pronouns in specific domain texts; the types of usable corpora (monolingual and/or bilingual) for examining the extraction of resolution rules have been changed depending on the type of NLP system using extracted resolution rules. The extraction processes of resolution rules in the environment are classified into five component tasks: (1) Zero Pronoun Identification, (2) Antecedent Annotation, (3) Rejection of Sentences Unsuitable for Rule Extraction, (4) Rule Extraction, and (5) Extracted Rule Application and Modification. An automatic process and/or a manual process with a user friendly human interface can be used to achieve each component task. This environment was implemented in the Japanese-to-English machine translation system, **ALT-J/E**, for Japanese zero pronoun resolution.

1 Introduction

In natural languages, elements that can be easily deduced by a reader are frequently omitted from expressions in texts (Kuno, 1978). This phenomenon causes considerable problems in NLP systems such as MT, text summarization and text retrieval. In particular, the subject and object are often omitted in Japanese, whereas they are normally obligatory in English[1]. In Japanese-to-English machine translation systems, therefore, it is necessary to identify case elements omitted from the original Japanese ("zero pronouns") for their accurate translation into English expressions.

Several algorithms have been proposed with regard to this problem (Kameyama, 1986; Yoshimoto, 1988; Walker et al., 1990; Dohsaka, 1994). However, it is not possible to apply these methods directly to a practical machine translation system because of their low precision of resolution and the large volume of knowledge required.

To overcome these kinds of problems, several methods have been proposed (Nakaiwa and Ikehara, 1992; Nakaiwa and Ikehara, 1995; Nakaiwa and Ikehara, 1996). The focus of these methods is on applications for a practical machine translation system with an unlimited translation target area.

With these methods, however, it is necessary to make resolution rules for zero pronouns by hand. Unfortunately, it takes a lot of time and effort for the experts of the NLP system to make these rules robust and with wide coverage. Furthermore, resolution rules often have to be made depending on the target domain of the documents, and this also requires the time-consuming labor of experts. Because of these problems, there is a need for an effective and efficient method of making resolution rules for zero pronouns.

Typical methods for this purpose include extracting resolution rules for zero pronouns from monolingual corpora (Nasukawa, 1996; Murata and Nagao, 1997), from bilingual corpora (Nakaiwa, 1997a; Nakaiwa, 1997b), and from monolingual corpora with tags for antecedents of zero pronouns (Aone and Bennett, 1995; Yamamoto and Sumita, 1998).

Monolingual corpora are relatively easy to collect. Methods using monolingual corpora, however, have difficulties in extracting resolution rules of zero pronouns whose referents are normally unexpressed in Japanese.

Methods using sentence-aligned bilingual corpora are better than those using monolingual corpora. This is particularly so with a bilingual corpus of dissimilar languages such as Japanese and English whose language families are so different and where the distributions of zero pronouns are also quite different. However, bilingual corpora are relatively difficult to collect,

[1] For example, there are 93 omitted case elements in 102 sentences in 30 newspaper articles which have to be explicitly translated into English.

especially sentence-aligned corpora.

With methods using monolingual corpora with antecedent tags, it is possible to efficiently make effective resolution rules by relying on the annotated information. However, there are only a few corpora with antecedent tags for zero pronouns. The standardization for annotating zero pronouns and their antecedents is still ongoing (Hasida, 2000). Consequently, in actual situations, analysts who want to make resolution rules for zero pronouns also have to laboriously annotate antecedent tags to zero pronouns in the corpus by hand, as previously mentioned. Therefore, an annotation tool for the antecedents of zero pronouns in the texts (Aone and Bennett, 1994) is needed for the effective addition of tags to zero pronouns.

To create resolution rules of zero pronouns in a text of a specific domain, we commonly use only monolingual corpora in the specific domain without antecedent tags for zero pronouns. Accordingly, analysts annotate tags to the antecedents of every zero pronoun in the corpus to make effective resolution rules. However, to accomplish this in machine translation, it is also possible to use bilingual corpora in the specific domain, such as a former version of a text that has already been translated or bilingual corpora used for translation memory systems. In this case, methods that automatically extract the resolution rules of zero pronouns from bilingual corpora (Nakaiwa, 1997a; Nakaiwa, 1997b) can be used. An automatic extraction process, however, cannot make perfect rule sets. Therefore, the automatically extracted rules have to be confirmed by human interaction before adding the rule set used in anaphora resolution in NLP systems if highly reliable rules such as domain-independent default rules are required. Furthermore, the human interaction must take into account the efficiency of acquiring resolution rules from both monolingual and bilingual corpora.

Considering these practical conditions for extracting the resolution rules of zero pronouns, this paper proposes a practical integrated tool capable of extracting rules for the anaphora resolution of zero pronouns from monolingual and/or bilingual corpora.

2 Component Tasks of Resolution Rule Extraction of Zero Pronouns

We classify the subtasks for extracting resolution rules from corpora into the following five component tasks: (1) Zero Pronoun Identification, (2) Antecedent Annotation, (3) Rejection of Sentences for Rule Extraction, (4) Rule Extraction, and (5) Extracted Rule Application and Modification.

2.1 Zero Pronoun Identification

The zero pronoun identification process identifies zero pronouns that must be resolved in an NLP system using extracted resolution rules. For example, Japanese, which is a free word-order language, often has no explicit cue helpful in determining obligatory case elements. Therefore, in this language, the identification of zero pronouns in the corpus is also important for extracting resolution rules. Furthermore, depending on the NLP system, the zero pronouns that must be resolved are different. For example, MT systems only need to resolve zero pronouns that must be explicitly translated into the target. In a Japanese sentence (1), the subject (*ga*-case) is not expressed in Japanese but becomes optional when translated into English, because it is possible to translate this by using the expression, "Zoos raise lions.".

(1) (*ϕ-ga*) *doubutsuen-de* *raion-o* *kau.*
ZOO-AT lion-OBJ keep
Zoos raise lions.

Therefore, in the zero pronoun identification process, the analysis results of the NLP system must be taken into account.

Zero pronoun identification in monolingual corpora only relies on the analysis results of the NLP system. In bilingual corpora, however, the translation equivalent of zero pronouns is also usable as a trigger for determining zero pronouns that must be resolved.

2.2 Antecedent Annotation

The antecedent annotation process identifies antecedents of zero pronouns that need to be resolved. In monolingual corpora, analysts must basically annotate antecedents of zero pronouns manually. However, even in the manual process, the following factors must be taken into account.

- Zero pronouns with the same syntactic and semantic features (such as modal expressions, the meaning of verbs, and conjunctions) around them in the corpus should be grouped and displayed at the same time when their antecedents are annotated.
  Zero pronouns with the same features tend to have the same type of antecedents because the features become key factors in

determining their antecedents. Therefore, analysts can judge antecedents for zero pronouns with the same features easily and efficiently.

- Antecedent candidates of zero pronouns should be easy to select from elements in the text or deictic elements outside the text.

  There are three types of possible antecedent candidates for each zero pronoun: candidates in the same sentence (intrasentential), candidates in another sentence in the text (intersentential), and candidates that are not explicitly expressed in the text (deictic). Intrasentential and intersentential antecedent candidates are actually expressed in the text. Their conditions in the resolution rules involve their syntactic positions and/or sentential relationships such as distance, rhetorical relation, and relative relation in the discourse structure (e.g., a candidate in the title of a section and a zero pronoun in a sentence in the section) (Nakaiwa and Ikehara, 1992; Nakaiwa and Ikehara, 1995). Therefore, by grouping intra- and intersentential candidates with the same syntactic position and sentential relationship in the text, and by showing the same types of candidates for zero pronouns at the same time, we can select the actual antecedent easily and efficiently. Among deictic antecedent candidates, the antecedents tend to be limited elements such as *writer/speaker* or *reader/hearer* (Nakaiwa and Ikehara, 1996). Therefore, listing the possible antecedent candidates before the annotation process and selecting the actual antecedent from the possible antecedent candidate list make the annotation process of deictic antecedents for zero pronouns much easier and more efficient.

In the case of bilingual corpora, in addition to the manual process used for monolingual corpora, the translation of a sentence with zero pronoun can be used to determine the antecedent of the zero pronoun. For example, in Japanese and English bilingual corpora, the subject and object are often omitted in Japanese, whereas they are normally obligatory in English. Therefore, by aligning zero pronouns in Japanese and their translation equivalents in English, antecedent of zero pronouns can be automatically identified (Nakaiwa, 1997b).

2.3 Rejection of Sentences Unsuitable for Rule Extraction

The following types of sentences with zero pronouns and/or antecedents in the corpus are not suitable as the source sentences for extracting rules.

(a) Sentences in which the analysis made errors in identifying the predicate, e.g., an adverbial expression, modal expression, or postpositional phrase as a predicate. This type of error identifies zero pronouns erroneously.

(b) Translation-equivalent sentences in bilingual corpora that were freely translated by a human. Here, it is very difficult to identify the translation equivalents of the zero pronouns within the translation-equivalent sentence in the automatic identification process.

The problematic sentences with zero pronouns and/or erroneous zero pronouns in type (a) have to be annotated as 'unsuitable' before extracting rules from sentences with zero pronouns in the corpus. The antecedents of zero pronouns in problematic sentences in type (b) have to be manually annotated even with bilingual corpora.

2.4 Rule Extraction

The rule extraction process extracts resolution rules of zero pronouns with antecedent tags in the corpus. In this process, syntactic and semantic features around zero pronouns and around their annotated antecedents are used as a condition in the resolution rules.

There are two way to extract rules:

(a) Automatic Extraction

In this process, resolution rules can be automatically extracted from zero pronouns with antecedent tags (Section 2.2) and syntactic and semantic features around zero pronouns and their annotated antecedents by a machine learning technique (Aone and Bennett, 1995; Yamamoto and Sumita, 1998) or by statistical processing (Nakaiwa, 1997a).

(a) Manual Extraction

In this process, zero pronouns are grouped depending on their syntactic position, their annotated antecedent, and the syntactic and semantic features around the zero pronouns. Resolution rules for the grouped zero pronouns are extracted by examining how many correct antecedents for the zero pronouns can be covered under the same features.

2.5 Extracted Rules Application and Modification

The extracted rules in section 2.4 are used by the NLP system in the resolution of zero pronouns in sentences in the corpus used for the rule extraction. Considering the results of the application of extracted rules for zero pronouns, we examine the suitability of rules for the corpus. If there are some problems in the resolution rules, the problematic rules and/or the priorities of the rules are modified.

The rule set with the modified rules is again used by the NLP system for the same corpus, and the suitability of rule modifications is checked in the same manner. The rule modification and re-application for zero pronouns within the corpus are iterated until reasonable rules are extracted.

3 Implemented Architecture for Extracting Resolution Rules of Zero Pronouns from Corpora

Considering the five components described in section 2, we have implemented an architecture for automatically and/or manually extracting resolution rules for Japanese zero pronouns from Japanese and English bilingual corpora and/or Japanese monolingual corpora. Figure 1 shows an overview of the system. In the first step, the Japanese and English sentences in the bilingual corpus and/or the Japanese sentences in the monolingual corpus are separately analyzed by Japanese and English parsers. In the next step, the antecedents of zero pronouns within the Japanese sentences in the corpus are identified automatically from Japanese and English analysis results in the bilingual corpus. From the monolingual corpus, however, only the Japanese analysis results with the syntactic position of zero pronouns are extracted. The Japanese analysis results with/without antecedent tags for zero pronouns are stored as 'Japanese Corpus with Antecedent Tags' as shown in the figure. Each zero pronoun in the corpus is manually examined in order to annotate the correct antecedent tags, if required. From the annotated information, resolution rules for zero pronouns are extracted manually or automatically. Manual extraction is preferable for acquiring reliable rules, but requires a high cost. In contrast, the automatic extraction process has a high possibility of extracting problematic rules. Such types of automatically extracted rules require human checks and modifications for extracting reliable rule sets.

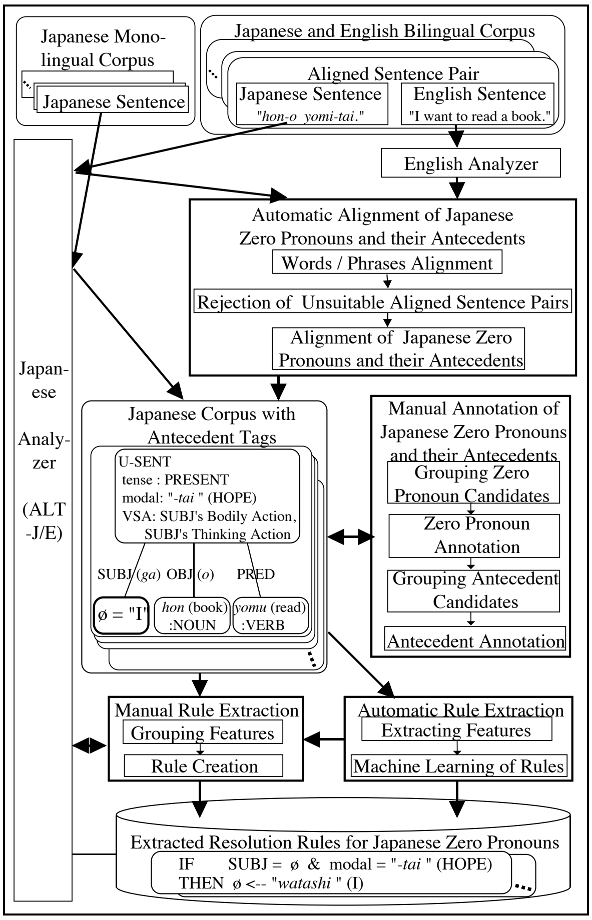

Japanese Mono-lingual Corpus
Japanese Sentence
Japanese and English Bilingual Corpus
Aligned Sentence Pair
Japanese Sentence
"hon-o yomi-tai."
English Sentence
"I want to read a book."
English Analyzer
Automatic Alignment of Japanese Zero Pronouns and their Antecedents
Words / Phrases Alignment
Rejection of Unsuitable Aligned Sentence Pairs
Alignment of Japanese Zero Pronouns and their Antecedents
Japan-ese Analy-zer (ALT -J/E)
Japanese Corpus with Antecedent Tags
U-SENT
tense : PRESENT
modal: "-tai " (HOPE)
VSA: SUBJ's Bodily Action, SUBJ's Thinking Action
SUBJ (ga)
OBJ (o)
PRED
ø = "I"
hon (book) :NOUN
yomu (read) :VERB
Manual Annotation of Japanese Zero Pronouns and their Antecedents
Grouping Zero Pronoun Candidates
Zero Pronoun Annotation
Grouping Antecedent Candidates
Antecedent Annotation
Manual Rule Extraction
Grouping Features
Rule Creation
Automatic Rule Extraction
Extracting Features
Machine Learning of Rules
Extracted Resolution Rules for Japanese Zero Pronouns
IF SUBJ = ø & modal = "-tai " (HOPE)
THEN ø <-- "watashi " (I)

Figure 1: Process for Extraction of Resolution Rules for Japanese Zero Pronouns

In the next step, the extracted resolution rules are used for the anaphora resolution of Japanese zero pronouns in the corpora by the Japanese analyzer. The same sentences in the monolingual and/or bilingual corpora are inputted to the system and resolution rules are again extracted and modified for the Japanese zero pronouns. These processes are repeated until the system cannot extract any more resolution rules for the Japanese zero pronouns in the corpora.

This method has been implemented in a Japanese-to-English machine translation system, **ALT-J/E** (Ikehara and et al., 1991). The system in Figure 1 can extract English translation equivalents of Japanese zero pronouns from aligned sentence pairs. Accordingly, the results can also be used to extract rules for translating Japanese zero pronouns into English in a Japanese-to-English machine translation system. For efficient human interaction in the manual process, we use the interface of a widely used WWW browser, such as Netscape Navigator or Internet Explorer.

In the following subsections, we describe the details of automatically and manually extracting resolution rules for the Japanese zero pronouns in the corpora.

3.1 Analysis of Japanese and English Sentences

Japanese sentences and English sentences in the corpora are analyzed in the following manner.

3.1.1 Analysis of Japanese Sentences

Japanese sentences are analyzed with the morphological, syntactic, and semantic analyzers of Japanese in **ALT-J/E** (Ikehara and et al., 1991). The syntactic and semantic structure of the Japanese sentence is first created. The Japanese structure is used for the automatic translation into English in **ALT-J/E**. The Japanese structure, therefore, includes the syntactic positions of the Japanese zero pronouns, which must be translated into English, and the semantic constraints for the Japanese zero pronouns forced by the verb within the Japanese sentence. When a zero pronoun is resolved by a rule, a determined antecedent and an ID of the applied rule for each zero pronoun are also annotated. This information is used to judge whether existing rules will resolve zero pronouns successfully, and which zero pronouns require further resolution rules.

For example, from the Japanese sentence in the aligned sentence pair (2) in Figure 1, the following syntactic and semantic structure is created.

(2)	(*ϕ-ga*)	*hon-o*	*yomi-tai*
		book-OBJ	read-WANT-TO
	I want to read a book.		

(3) Syntactic and Semantic Structure of Japanese Sentence (2)

$$\begin{bmatrix} \texttt{U\_SENT-1} & \\ \texttt{Tense} & \text{PRESENT, PERFECTIVE ASPECT} \\ \texttt{Modal} & \textit{tai}\ \text{(HOPE)} \\ \texttt{VSA} & \text{SUBJECT'S HUMAN ACTION,} \\ & \text{SUBJECT'S THINKING ACTION} \\ \texttt{PRED-1} & \begin{bmatrix} \texttt{main verb} & \textit{yomu}\ \text{``}\textit{read}\text{''} \end{bmatrix} \\ \texttt{CASE-1} & \begin{bmatrix} \texttt{case rel.} & \text{OBJECT ``}\textit{o}\text{''} \\ \texttt{NP-1} & \textit{hon}\ \text{``}\textit{book}\text{''} \end{bmatrix} \\ \texttt{CASE-2} & \begin{bmatrix} \texttt{case rel.} & \text{SUBJ} \\ \texttt{NP-2} & \phi\text{-1} \\ \texttt{semantic constraints} & \text{HUMAN} \end{bmatrix} \end{bmatrix}$$

3.1.2 Analysis of English Sentences

English sentences are analyzed by Brill's English Tagger (Brill, 1992) and the Link Grammar Parser (Sleator and Temperley, 1991). Next, the syntactic structure is converted into a partial syntactic structure, which is similar to the internal English structure of **ALT-J/E**.

For example, from the English sentence in aligned sentence pair (2), the following partial syntactic structure is created.

(4) Partial syntactic structure of an English Sentence (2)

$$\begin{bmatrix} \texttt{U\_SENT-1} & \\ \texttt{PRED-1} & \begin{bmatrix} \text{``}\textit{want}\text{''} & \text{VERB, SING., PRESENT.} \\ \text{``}\textit{to}\text{''} & \text{TO} \\ \text{``}\textit{read}\text{''} & \text{VERB, BASE FORM} \end{bmatrix} \\ \texttt{CASE-1} & \begin{bmatrix} \texttt{case rel.} & \text{SUBJECT} \\ \texttt{NP-1} & \begin{bmatrix} \text{``}\textit{I}\text{' : PERSONAL PRONOUN} \end{bmatrix} \end{bmatrix} \\ \texttt{CASE-2} & \begin{bmatrix} \texttt{case rel.} & \text{DIRECT OBJECT} \\ \texttt{NP-2} & \begin{bmatrix} \text{``}\textit{a}\text{''} & \text{DETERMINER} \\ \text{``}\textit{book}\text{''} & \text{NOUN,} \\ & \text{SINGULAR OR MASS} \end{bmatrix} \end{bmatrix} \end{bmatrix}$$

3.2 Automatic Alignment of Japanese Zero Pronouns and their Antecedents[2]

From the analysis results of Japanese and English aligned sentence pairs, the system extracts pairs of Japanese words/phrases and their English equivalent words/phrases by comparing the two structures. Then, based on the discussion in section 2.3, aligned sentence pairs not suitable for the extraction of resolution rules for Japanese zero pronouns are automatically identified if any of the following conditions are met.

- There is a difference between the number of clauses whose Japanese verb is not aligned with an English noun, within the Japanese analysis result, and the number of clauses within the English analysis result.
- The MT system fails to translate some words.
- The English Parser is unable to make a full syntactic structure.

Next, the Japanese zero pronouns in the Japanese sentences and the translation equivalents of their antecedents in English sentences are extracted using 10 hand-developed alignment rules.

For example, from the zero pronoun in the *ga*-case (subject) in the Japanese sentence in

[2] This process is implemented by using the alignment method proposed by Nakaiwa (Nakaiwa, 1997b).

aligned sentence pair (2), its antecedent is automatically determined as the subject in the English sentence ("*I*"), as shown in the 'Japanese Corpus with Antecedent Tags' block in Figure 1.

3.3 Manual Annotation of Japanese Zero Pronouns and their Antecedents

With Japanese monolingual corpora, an analyst who wants to make resolution rules for Japanese zero pronouns in the corpora must annotate their antecedents by hand. To achieve an efficient annotation process, we have developed a tool for annotating antecedents of Japanese zero pronouns in Japanese sentences within the corpora. This process uses the analysis results of Japanese sentences (section 3.1.1). The details of this process are described in the following sections.

3.3.1 Identifying Zero Pronouns

According to the results of the Japanese analysis in a Japanese-to-English MT system, the zero pronouns that must be explicitly translated in English are explicitly annotated in the syntactic and semantic structure of the inputted Japanese sentences (e.g., example (3)). However, as we discussed in section 2.3, the sentence analysis error causes erroneous zero pronouns. Therefore, the analyst must annotate whether the zero pronoun candidates in the Japanese analysis result are actually zero pronouns or not. For efficiency, the Japanese analysis results are grouped based on whether the same features are around zero pronoun candidates as follows.

(1) syntactic position of zero pronoun candidates (e.g., *ga-case* (Subject), *o-case* (Direct Object)).

(2) syntactic and semantic structure around zero pronoun candidates (e.g., the types of conjunctions, verbal semantic attributes, and the types of modal expressions in unit sentences with zero pronouns candidates).

Figure 2 shows an example of the display after grouping zero pronoun candidates according to their syntactic positions. As shown in the figure, N1 (*ga*-case) is the most common syntactic position of zero pronouns in the corpus (866 instances in 724 sentences), and N2 (*o*-case) is the next most common (125 instances in 116 sentences).

3.3.2 Annotating Antecedents of Zero Pronouns

After identifying zero pronouns in the Japanese sentences, an antecedent for each zero pronoun is annotated. As with the process of zero pronoun identification, zero pronouns are grouped based on the presence of the same characteristics around the zero pronouns. To efficiently annotate the antecedents of zero pronouns, we also group intrasentential and intersentential anaphora candidates according to the characteristics around the candidates as follows:

(1) syntactic position of an antecedent candidate

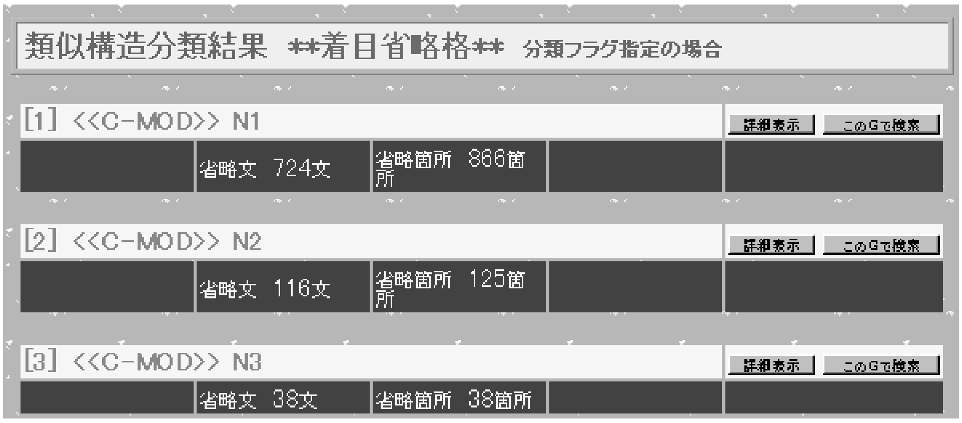
(2) syntactic and semantic structures around an antecedent candidate

(3) syntactic relationship between a unit sentence with a zero pronoun and a unit sentence with an antecedent candidate in the same sentence (intrasentential) (e.g., a unit sentence with a zero pronoun is directly connected to another unit sentence by a conjunction)

(4) discourse structural relationship (or distances) between a sentence with a zero pronoun and a sentence with an antecedent candidate (intersentential) (e.g., a sentence with a zero pronoun is the next sentence following a sentence with an antecedent candidate)

For (3), the syntactic structures of unit sentences with zero pronouns are classified, and the sentences with the same types of syntactic structures are examined. For (4), typical antecedent candidate relationships for the target corpus are selected in advance. For example, in newspaper articles, the first sentence of an article often contains the antecedent of a zero pronoun in another sentence in the article (Nakaiwa and Ikehara, 1992). The relationship between sentences should be selected depending on the target domain to achieve an efficient annotating process. An analyst annotates deictic antecedents of zero pronouns by first selecting typical antecedent candidates such as "I/we", "you", or "it" in advance and then selecting the diectic antecedent of a zero pronoun from them. After an antecedent for a zero pronoun is annotated, other antecedent candidates for the zero pronoun are displayed as "negative candidates" in the display of the grouping result.

By grouping antecedent candidates having the same characteristics, analysts can efficiently



類似構造分類結果 **着目省略格** 分類フラグ指定の場合

[1] <<C-MOD>> N1

省略文 724文 省略箇所 866箇所

[2] <<C-MOD>> N2

省略文 116文 省略箇所 125箇所

[3] <<C-MOD>> N3

省略文 38文 省略箇所 38箇所

Figure 2: Display of Grouping Result of Zero Pronouns Candidates according to their Syntactic Positions

annotate antecedents of zero pronouns by referring to the antecedent candidates within the same type of context. Furthermore, by annotating antecedents from the context with high frequency to low frequency, an analyst can efficiently annotate antecedents in the early stage of the annotating process.

3.4 Automatic Extraction of Resolution Rules

Syntactic and semantic features around zero pronouns and their antecedents are extracted from Japanese sentences with Japanese analysis results and with tags for zero pronouns and their antecedents. The following features, the effects of which were discussed in Nakaiwa (1992;1995;1996) are used as conditions of extracted resolution rules.

- verbal semantic attributes (Nakaiwa et al., 1994)
- type of modal expression (Kawai, 1987)
- type of conjunction between a unit sentence with a zero pronoun and a unit sentence with its antecedent
- syntactic relationship between a unit sentence with a zero pronoun and a unit sentence with its antecedent (intrasentential)
- discourse structural relationship (or distance) between a sentence with a zero pronoun and a sentence with its antecedent (intersentential)

Rules are automatically extracted by a decision tree learning program, C5.0 (Quinlan., 1998). Extracted rules are converted to the rule format used in **ALT-J/E**.

3.5 Manual Extraction of Resolution Rules

For the extraction of more reliable resolution rules with human interaction, a manual rule extraction process from Japanese sentences using Japanese analysis results and tags for zero pronouns and their antecedents is implemented in the system. In the same manner as in section 3.3, the five types of features around zero pronouns and their antecedents used in section 3.4 are grouped and sorted by the frequencies of the grouped items. Therefore, wide coverage rules are efficiently extracted in the early stage of the extraction process. This is also effective for rule extraction by taking into account zero pronouns with the same types of context. The reliability of the extracted rules is also examined in this stage by calculating the number of applied zero pronouns for each rule and the number of successfully resolved zero pronouns by referring to antecedent tags for zero pronouns. Before the extracted rules are added to the rule set used in **ALT-J/E**, inclusion relationships between rules are examined and the priorities of extracted rules within the rule set are set.

4 Preliminary Evaluation

The performance of the automatic extraction process from aligned sentence pairs has been reported in (Nakaiwa, 1997a; Nakaiwa, 1997b). According to the evaluation result on the automatic alignment of Japanese zero pronouns and the English equivalents of their antecedents, 98.4% of all pairs were automatically aligned correctly in the training data and 94% of all pairs in unseen test data. Furthermore, according to their evaluation of extracted rules for zero pronouns with deictic references, those rules created automatically from sentence pairs correctly resolved 99.0% of the zero pronouns in the training data and 85.0% of the zero pronouns in an unseen test data. Therefore, we only evaluate the manual extraction process from Japanese monolingual corpora. The effectiveness of the proposed method of manual rule extraction highly relies on their grouping function. Therefore, in this evaluation, we examine the effectiveness of the manual rule extraction with or without the grouping function.

4.1 Evaluation Method of Manual Rule Extraction

The effectiveness and efficiency of manual rule extraction is examined by extracting rules from 3719 Japanese sentences in a test set for evaluating Japanese-to-English MT system (Ikehara et al., 1994). An analyst who is an expert of zero pronoun resolution in **ALT-J/E**, extracts resolution rules using the implemented system in section 3, which is installed in SUN Sparc Enterprise 3000, in the following manner.

A. Extraction using the Grouping Function
The grouping, annotation and extraction are conducted as follows.

Step 1 Zero pronoun candidates are grouped according to their syntactic positions; the candidates in the most common syntactic position, N1 (*ga*-case) are selected: 866 instances in 724 sentences (Figure 2)

Step 2 Selected candidates are grouped again according to their syntactic structure; the candidates in the most common syntactic structure, where a unit sentence with a zero pronoun is directly connected with another unit sentence by a conjunction, are selected: 315 instances

Step 3 Zero pronouns are annotated for the selected candidates: 285 out of 315 instances

Step 4 Antecedents of selected zero pronouns are annotated for 285 zero pronouns after grouping the type of conjunctions.

Step 5 Five rules are extracted from 285 zero pronouns and the required time for making the rules is recorded.

B. Extraction without using the Grouping Function
Rules are extracted from sentences with zero pronoun candidates one by one without using the grouping function in the time it takes to make the five rules in test A.

The results of two tests are compared by examining how fast the antecedents of zero pronouns can be efficiently annotated and how many zero pronouns can be successfully resolved by using extracted rules.

4.2 Evaluation Result

Table 1 shows the results of the evaluation. As shown in this table, zero pronouns and their antecedents were efficiently annotated in test A (1.1 min/item and 1.7 min/item using the grouping function (test A), and 2.5 min/item and 2.0 min/item without using the grouping function (test B), respectively). Furthermore, the rule extraction time and its application and evaluation time were also shorter in test A than in test B (1.1 min/item and 2.2 min/item in test A, and 6.0 min/item and 10.0 min./item in test B, respectively). This result indicates that grouping results with annotated information is helpful for making rules with wide coverage by taking into account the estimated result of an extracting rule for zero pronouns that will be applied. Regarding the quality of extracted rules, test B extracted better rules than test A (93 % in test A and 100 % in test B). However, the five problematic zero pronouns in test A were already noticed by the analyst at the rule evaluation step. Therefore, new rules for the zero pronouns with a detailed condition will be extracted easily by referring to this result.

Table 1: Required Time and Accuracy of Manually Extracted Rules (required time per zero pronoun shown in parentheses)

Grouping Function		used (A)	unused (B)
# of Extracted Rules		5	51
Required Time [min]	Zero Pron. Identification	332 (1.1)	128 (2.5)
	Antecedent Identification	482 (1.7)	102 (2.0)
	Rule Extraction	77 (1.1)	306 (6.0)
	Rule Application and Evaluation	158 (2.2)	510 (10.0)
	Total	1049 (6.0)	1046 (20.5)
# of Applied Zero Pron.		72	51
# of Correctly Resolved Zero Pron.		67 (93%)	51 (100%)

5 Conclusions

This paper proposed a practical integrated tool for extracting rules for the anaphora resolution of zero pronouns from monolingual and/or bilingual corpora. According to the preliminary evaluation of the manual rule extraction process, antecedent tags for zero pronouns can be efficiently annotated and rules are efficiently extracted from Japanese monolingual corpora by using the tool's grouping function. In the future, we will examine the effectiveness of the proposed method for both monolingual and bilingual corpora. We will also examine the most effective combined strategies for the extraction of resolution rules by using both automatic and manual processes.

References

Chinatsu Aone and Scott W. Bennett. 1994. Discourse tagging tool and discourse-tagged multilingual corpora. In *Proc. of the Internationl Workshop on Sharable Natural Language Resources*, pages 71–77.

Chinatsu Aone and Scott W. Bennett. 1995. Automated acquisition of anaphora resolution strategies. In *Working Notes of AAAI Spring Symposium Series, Empirical Methods in Discourse Interpretation and Generation*, pages 1–7.

Eric Brill. 1992. A simple rule-based part of speech tagger. In *Proc. of ANLP-92*, pages 152–155.

Kohji Dohsaka. 1994. Identifying the referents of Japanese zero-pronouns based on pragmatic condition interpretation. *Transaction of IPSJ*, 35(10):34–40. In Japanese.

Koiti Hasida. 2000. Global document annotation (GDA). `http://www.etl.go.jp/etl/nl/GDA/`.

Satoru Ikehara and Satoshi Shirai et al. 1991. Toward MT system without pre-editing – effects of new methods in **ALT-J/E** –. In *Proc. of MT Summit III*, pages 101–106. (`http://xxx.lanl.gov/abs/cmp-lg/9510008`).

Satoru Ikehara, Satoshi Shirai, and Kentaro Ogura. 1994. Criteria for evaluating the linguistic quality of Japanese-to-English machine translation. *Journal of JSAI*, 9(5):569–579.

Megumi Kameyama. 1986. A property-sharing constraint in centering. In *Proc. of the 24th Annual Meeting of ACL*, pages 200–206.

Atsuo Kawai. 1987. Modality, tense and aspect in Japanese-to-English translation system ALT-J/E. In *Proc. of the 34th Annual Conv. of IPSJ*, pages 1245–1246. In Japanese.

Susumu Kuno. 1978. *Danwa no Bunpoo.* Taishukan Publ. Co., Tokyo, Japan. In Japanese.

Masaaki Murata and Makoto Nagao. 1997. An estimation of referents of pronouns in Japanese sentences using examples and surface expressions. *Journal of Natural Language Processing*, 4(1):87–109.

Hiromi Nakaiwa and Satoru Ikehara. 1992. Zero pronoun resolution in a Japanese-to-English machine translation system by using verbal semantic attributes. In *Proc. of ANLP-92*, pages 201–208.

Hiromi Nakaiwa and Satoru Ikehara. 1995. Intrasentential resolution of Japanese zero pronouns in a machine translation system using semantic and pragmatic constraints. In *Proc. of TMI-95*, pages 96–105.

Hiromi Nakaiwa and Satoru Ikehara. 1996. Anaphora resolution of Japanese zero pronouns with deictic reference. In *Proc. of COLING-96*, pages 812–817.

Hiromi Nakaiwa, Akio Yokoo, and Satoru Ikehara. 1994. A system of verbal semantic attributes focused on the syntactic correspondence between Japanese and English. In *Proc. of COLING-94*, pages 672–678.

Hiromi Nakaiwa. 1997a. Automatic extraction of rules for anaphora resolution of Japanese zero pronouns from aligned sentence pairs. In *Proc. of ACL-97/EACL-97 Workshop on Operational Factors in Practical, Robust, Anaphora Resolution for Unrestricted Texts*, pages 22–29. ACL.

Hiromi Nakaiwa. 1997b. Automatic identification of zero pronouns and their antecedents within aligned sentence pairs. In *Proc. of the 5th WVLC*, pages 127–141.

Tetsuya Nasukawa. 1996. Full-text processing: Improving a practical NLP system based on surface information within the context. In *Proc. of COLING-96*, pages 824–829.

J. Ross Quinlan. 1998. `http://www.rulequest.com/`.

Daniel Sleator and Davy Temperley. 1991. Parsing English with a link grammar. *Carnegie Mellon University Computer Science Technical Report*, pages CMU–CS–91–196.

Marilyn Walker, Masayo Iida, and Sharon Cote. 1990. Centering in Japanese discourse. In *Proc. of COLING-90*, pages 1–6.

Kazuhide Yamamoto and Eiichiro Sumita. 1998. Feasibility study for ellipsis resolution in dialogues by machine-learning technique. In *Proc. of COLING-ACL-98*, pages 1428–1434.

Kei Yoshimoto. 1988. Identifying zero pronouns in Japanese dialogue. In *Proc. of COLING-88*, pages 779–784.