Automatic Identification of Zero Pronouns and their Antecedents within Aligned Sentence Pairs

Hiromi Nakaiwa
NTT Communication Science Laboratories
1-1 Hikarinooka, Yokosuka-shi,
Kanagawa-ken, 239 JAPAN
nakaiwa@cslab.kecl.ntt.co.jp

Abstract

This paper proposes a method to identify zero pronouns within a Japanese sentence and their antecedent equivalents within the corresponding English sentence from aligned sentence pairs. The method focuses on the characteristics of Japanese and English, in two languages from different families and in which distribution of zero pronouns is very different. In this method, the Japanese sentence and English translation within the Japanese and English aligned sentence pairs are analyzed. Then, the pairs of Japanese word/phrase and their English equivalent word/phrase are identified from each aligned sentence pair. Next, zero pronouns within a Japanese sentence are identified by using the syntactic and semantic structure of the Japanese sentence and their antecedents within the English sentence are identified by using the characteristics of anaphoric and deictic expressions in English. This method was implemented using the Japanese-to-English machine translation system, ALT-J/E for the analysis of Japanese sentences and Brill's tagger for the analysis of the English sentences. According to my evaluation, for 554 zero pronouns in a sentence set for the evaluation of Japanese-to-English machine translation systems, 91.5% of the pairs of zero pronouns in the Japanese sentences and their antecedents in the English translations were automatically identified correctly.

1 Introduction

1.1 Motivation

In natural languages, elements that can be easily deduced by the reader are frequently omitted from expressions in texts (Kuno, 1978). This phenomenon causes considerable problems in natural language processing systems. For example, in a machine translation system, the system needs to recognize those elements which are not present in the source language, but may become mandatory elements in the target language. In particular, the subject and object are often omitted in Japanese; whereas they are normally obligatory in English. Thus, in Japanese-to-English machine translation systems, it is necessary to identify case elements omitted from the original Japanese ("zero pronouns") for their translation into English expressions.

Several algorithms have been proposed with regard to this problem (Kameyama, 1986; Walker et al., 1990; Yoshimoto, 1988; Dousaka, 1994). When considering the application of these methods to a practical machine translation system for which the translation target area can not be limited, it is not possible to apply them directly, both because their precision of resolution is low as they only use limited information, and because the volume of knowledge that must be prepared beforehand is so large.

To overcome these kinds of problems, several methods to resolve zero pronouns which consider applications for a practical machine translation system with an unlimited translation target area, have been proposed (Nakaiwa and Ikehara, 1992; Nakaiwa and Ikehara, 1995; Nakaiwa and Ikehara, 1996). These methods use categorized semantic and pragmatic constraints such as verbal semantic attributes (Nakaiwa et al., 1994) and types of modal expressions and conjunctions as a condition for anaphora resolution of zero pronouns.

But, with these methods it is necessary to make resolution rules for zero pronouns by hand. To make robust rules, with wide coverage, takes a lot of time and labor. Analysts who make these resolution rules must be familiar with the NLP system itself. Furthermore, the types of zero pronouns change depending on the types of documents which must be analyzed. So, resolution rules must be made depending on the target domain of the documents. But, it is very difficult to make rules for every domain because of the time consuming labor and the need for expertise. Because of these problems, a method to make resolution rules of zero pronouns effectively and efficiently is needed.

In order to acquire resolution rules for a NLP system effectively and efficiently, various methods have been proposed. One typical method for this purpose is to use a corpus for extracting resolution rules by analyzing each sentence in the corpus. With regard to the automatic extraction of resolution rules for zero pronouns, several methods have been proposed (Murata and Nagao, 1997; Nasukawa, 1996). But these methods use monolingual corpora and they find it difficult to extract resolution rules of zero pronouns whose referents are normally unexpressed in Japanese. Furthermore, rules can only be made when similar expressions to those containing the zero pronouns are found in the corpus.

It seems that a bilingual corpus consisting of sentence pairs, with an original in one language and a translation, is better than a monolingual corpus for the purpose of acquiring resolution rules of zero pronouns. This is particularly so with a bilingual corpus of Japanese and English whose language families are so different and in which the distribution of zero pronouns is also very different. This combination is more useful than the bilingual corpora of similar languages.

The technique for acquiring various kinds of rules such as translation rules, grammar rules, dictionary entries and so on from bilingual corpora needs to include several kinds of sub-techniques; identification of aligned sentence pairs which consist of pairs of one language sentence and translation equivalents of the sentence (sentence alignment); identification of equivalent words/phrases pairs from aligned sentence pairs (word alignment); and extraction of rules such as translation rules, grammar rules, dictionary entries and so on from identified aligned sentence pairs and equivalent word/phrase pairs.

Several methods have been proposed with regard to aligning sentences (Brown et al., 1991; Gale and Church, 1991; Haruno and Yamazaki, 1996; Kay and Roscheisen, 1993), aligning words (Church, 1993; Kupiec, 1993; Matsumoto et al., 1993; Wu, 1995; Yamada et al., 1996) and acquiring rules from bilingual corpora (Dagan et al., 1991; Dagan and Church, 1994; Fung and Church, 1994; Tanaka, 1994; Yamada et al., 1995). From the point of view of the extraction of resolution rules of zero pronouns, a technique to identify zero pronouns in a sentence in one language and their antecedents in a translation from aligned sentence pairs is needed. But there is currently no method to identify zero pronouns and their antecedents automatically within bilingual corpora.

1.2 A Method for Extraction of Resolution Rules for Japanese Zero Pronouns

This section describes the overall design of a method for automatically extracting resolution rules for Japanese zero pronouns from Japanese and English aligned sentence pairs.[1] Figure 1 shows an overview of the system. As shown in this figure, the aligned sentence pairs, consisting of a Japanese sentence and its English translation, are analyzed individually by Japanese and English syntactic and semantic parsers. Next, the system identifies the pairs of Japanese word/phrase and their English equivalent word/phrase, by comparing these two structures, based on the Japanese structure and the English structure which are created by the Japanese and English parsers. Then, Japanese zero pronouns in the Japanese sentence and translation equivalents of their antecedents in the English sentence are identified. By using these results, based on the Japanese structure, the resolution rules for Japanese zero pronouns within Japanese sentences are created. In the next step, the resolution rules are used for the semantic and pragmatic analysis of the Japanese sentence by the Japanese parser within the whole rule extraction system. The same Japanese and English aligned sentence pairs are inputted to the system and resolution rules of Japanese zero pronouns are again extracted. These processes are repeated until the system cannot extract any more rules for Japanese zero pronouns resolution from the aligned sentence pairs.

In this paper, I describe the core method to identify zero pronouns within a Japanese sentence and translation equivalents of their antecedents within a corresponding English sentence, shown by the shaded area in Figure 1.

2 Appearance of Zero Pronouns and Their Antecedents within Japanese and English Aligned Sentence Pairs

In order to understand the distribution of zero pronouns with antecedents within Japanese and English aligned sentence pairs, this section examines which zero pronouns in Japanese must be explicitly translated into English and where their antecedents appear in English, using a test set designed to evaluate the performance of Japanese-to-English machine translation systems (Ikehara et al., 1994). The test set (3718 sentences) has many examples of zero pronouns with intrasentential and deictic references.[2] The sentence set was created to test the coverage of expressions that can be translated by Japanese-to-English MT systems based on the varieties of Japanese expressions and the differences between Japanese and English. The sentence set has approximately 500 kinds of test items. Each sentence has a manual translation, and almost all of the sentences can be translated without contextual information (3704 sentences out of 3718 sentences). A MT system can be evaluated by comparing its output to the equivalent manual translation. Each sentence is expressed in natural Japanese and the sentence set covers many different expressions. The average length of Japanese sentence is 16.39 Japanese characters and the average length of English translation is 9.22 words.

This is an example of a zero pronoun in Japanese whose referent is expressed in the English equivalent.

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

[1] For details of this step, please refer to Nakaiwa (1997).

[2] 'intrasentential' means that the antecedent of Japanese zero pronoun is explicitly expressed within the same Japanese sentence. 'deictic' means that the antecedent of Japanese zero pronouns is not explicitly expressed within the Japanese text.

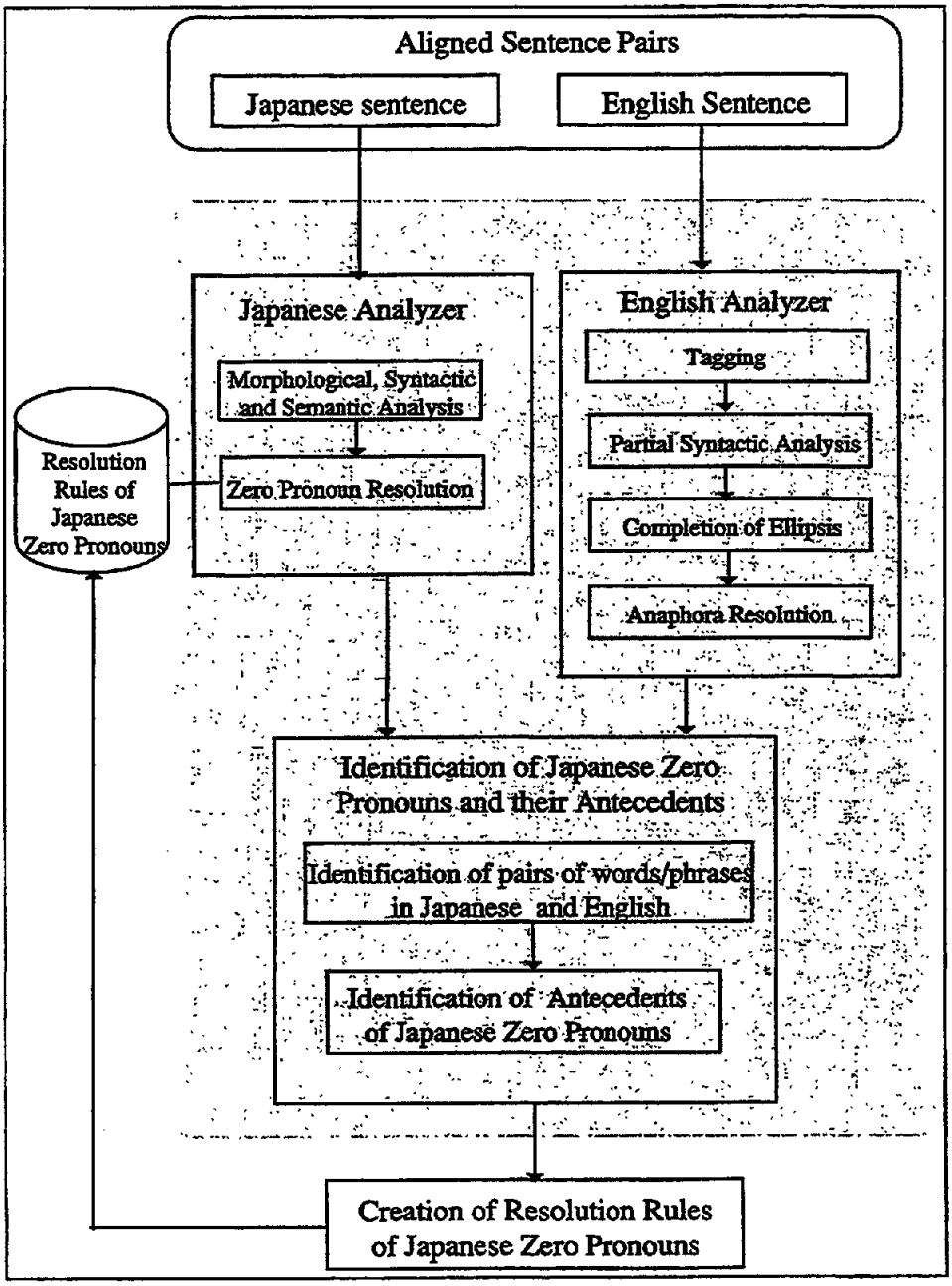

Figure 1: System of Automatic Extraction of Resolution Rules for Japanese Zero Pronouns

In this expression, the Japanese sentence contains the modal expression *tai* which indicates HOPE. This modal expression causes the default referent of the subject zero pronoun to be the "writer" or "speaker" which is translated as "I" in English.

The results of the examination of zero pronouns and their referential elements in the test set are shown in Table 1. There were a total of 554 zero pronouns in 437 sentences. The location of the referential elements can be divided into 2 kinds: those expressed in the same Japanese sentence, and those not expressed in the same Japanese sentence. The latter were further classified into 7 kinds.

- The zero pronoun is not translated because the passive voice is used.
- The referent is the writer or speaker, "I"; or a group, "we".
- The referent is the reader or hearer, "you".
- The referent is human but it is not known who the human is.
- The zero pronoun should be translated as "it".
- The zero pronoun should be translated as "this".
- The referent is another specific element.

According to this study of the functional test sentence set, in 150 out of 554 instances (27%) the antecedents were expressed within the Japanese sentence and in 404 out of 554 instances (73%) the referent was not expressed in the Japanese sentence. Zero pronouns could be left unexpressed by converting the translation to the passive voice in 157 instances (28%). The other zero pronouns, 247 instances (45%), referred to referents that did not appear in the Japanese sentence but did appear in the English translation. This result shows that aligned sentence pairs will be effective for identifying zero pronouns and their antecedents automatically by determining zero pronouns in Japanese and their antecedents in English.

According to a further examination of the English equivalents of Japanese zero pronouns in the sentence set in Table 1, depending on the types of referential elements, the style of the English equivalents are different. These characteristics can be summarized in the following:

1. Deictic referents in English (247 instances)
   These elements are often translated as personal pronouns such as "I" or "you" or indefinite "one".

2. Anaphoric antecedents in English (150 instances; intrasentential)
   These elements are often translated as personal pronouns, demonstratives, such as "this", definite noun phrases, such as noun phrases with definite articles (e.g. "the company") or anaphoric "one".

English expressions of these two types are preferred candidates for translation equivalents of Japanese zero pronouns.

Table 1: Distribution of zero pronouns and their referential elements

Loc. of zero pron.	Loc. of 'referential elements'												Total
	Intrasentential					Deictic							
	ha	*ga*	*o*	*ni*	misc	Psve	I we	you	HUMAN	it	this	misc	
ha	2	0	0	0	0	3	1	2	0	1	0	0	9
ga	113	12	1	1	8	152	88	25	31	57	3	0	491
o	4	0	6	1	0	0	0	0	2	19	1	0	33
ni	1	0	0	0	0	2	3	8	0	1	0	2	17
no	0	0	1	0	0	0	1	0	0	2	0	0	4
Total	150					404							554

3 A Method for Identification of Zero Pronouns and their Antecedents within Japanese and English Aligned Sentence Pairs

This section describes a method for automatically identifying zero pronouns within Japanese sentences, and translation equivalents of their antecedents within English translation equivalents. The shaded part of Figure 1 shows an overview of the system. The system consists of the following three parts; the analysis of the Japanese sentence; the analysis of the English sentence; and the identification of Japanese zero pronouns in Japanese sentences and their antecedents in English equivalent sentences. The next three subsections describe the details of these three process.

3.1 Analysis of Japanese Sentences

The Japanese sentences are analyzed using the morphological, syntactic and semantic analyzers of Japanese in ALT-J/E (Ikehara et al., 1991). The analysis of the Japanese sentences consists of the following steps.

1) **Morphological analysis of Japanese sentences**
In this step, Japanese sentence is segmented into words and each word is tagged with its part of speech.

2) **Syntactic and semantic analysis of Japanese sentences**
In this step, using each word within the Japanese sentence and the tagged information for each word, the syntactic and semantic structure of the Japanese sentence is created. Because the Japanese syntactic and semantic structure is used for the automatic translation to English in ALT-J/E, the Japanese structure contains information about the syntactic position of Japanese zero pronouns which must be translated into English and also contains the semantic constraints for the Japanese zero pronouns forced by the meaning of the verb within the Japanese sentence. For the semantic constraints, we used our original semantic attribute system(Ogura et al., 1993). The semantic attribute system has a semantic concept hierarchy which represents "is-a" relationships and "has-a" relationships, with some 2,800 attributes (12-level tree structure) for common nouns, and some 200 attributes (9-level tree structure) for proper nouns.

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

(2) Syntactic and Semantic Structure of Japanese Sentence (1)

```
S  :  u-sent-1
      tense    :  PRESENT , PERFECTIVE ASPECT
      modal    :  tai (HOPE)
      VSA      :  SUBJECT'S HUMAN ACTION, SUBJECT'S THINKING ACTION
   |- PRED     :  pred-1
   |             main verb : yomu (READ)
   |- CASE     :  case-1
   |             case relation : OBJECTIVE CASE
   |             particle : wo
   |    |- NP  :  np-1
   |         |-    N : hon (BOOK)
   |- CASE     :  case-2
                   case relation : SUBJECT
        |- NP   :  φ-1 (Semantic Constraints : HUMAN)
```

3.2 Analysis of English Sentences

The English sentences are analyzed using Brill's English tagger (Brill, 1992) and partially parsed using the tagged information for each English word. The analysis of English sentences consists of the following steps.

1) **Tagging for each word in the English sentence**
   Using Brill's English tagger, each English word in the English sentence is tagged with the appropriate part of speech.

2) **Partial syntactic analysis of the English sentence**
   In this step, first of all, noun phrases and the predicate part within the English sentence are identified, using the tagged information for each English word by step 1. The predicate part means the consecutive words which consist of verb and modal, tense and aspect. Each predicate part is distinguished by its voice. Next, the subject of the predicate part and the direct object of the predicate part are determined from noun phrases within the English sentence. The subject is simply taken to be the preceding adjacent noun phrase of the predicate part and the direct object the next adjacent noun phrase.[3]

3) **Completion of ellipsis within the English sentence**
   Ellipses within the English sentence are identified and are completed by using the syntactic structure. At present, only ellipsed coordinated subjects are treated. In practice, the following rules are used for the identification of ellipsis and completed elements.

   > **Determination of Ellipsis** When a predicate part in the English sentence is in the active voice and the adjacent preceding element is not a noun phrase but a coordinating conjunction, the subject of the predicate part is determined to be ellipsed.
   > **Determination of Omitted Element** When the subject of a predicate part is determined to be ellipsed and the subject of the preceding predicate part for the

[3] In this step, it is possible to use an English parser as the syntactic analyzer. But, normally, English parsers make many possible syntactic structures which are often incorrect. So, when an English parser is used for the English syntactic analysis for this step, the system must select the best structure from the candidates. Furthermore, as shown in the following step, the partial syntactic structures are all that is needed for the identification of antecedents of Japanese zero pronouns within the English translation. So, to keep the robustness, I only use Brill's English tagger and partial syntactic parser.

ellipsis exists or is ellipsed and their omitted element has already been determined, the subject noun phrase is determined as the omitted element for the ellipsis.

4) **Anaphora resolution of anaphoric expressions within the English sentence**
The anaphora resolution of anaphoric expression such as pronouns and definite noun phrases, is conducted using the partial syntactic structure of the English sentence. According to this step, even if the antecedents of zero pronouns in Japanese are anaphoric expressions such as pronouns and definite noun phrases, the anaphora resolution process determines the antecedents of anaphoric expressions in English and the overall system can determine the intersentential and intrasentential resolution rules of Japanese zero pronouns by using identified pairs of the antecedents of anaphoric expressions in English and their Japanese equivalents.[4]

For example from the English sentence in aligned sentence pair (1), the following partial syntactic structure is created.

(3) Partial syntactic structure of an English Sentence (1)

```
S  :  u-sent-1
   |- PRED   :  pred-1
   |            "want" : VERB, NON-3RD PS. SING. PRESENT.
   |            "to" : TO
   |            "read" : VERB, BASE FORM
   |- CASE   :  case-1
   |            case relation : SUBJECT
   |    |- NP :  np-1
   |         |- "I" : PERSONAL PRONOUN
   |- CASE   :  case-2
   |            case relation : DIRECT OBJECT
        |- NP :  np-2
             |- "a" : DETERMINER
                "book" : NOUN, SINGULAR OR MASS
```

3.3 Identification of Japanese Zero Pronouns and Their Antecedents

The method to identify Japanese zero pronouns and their English equivalents consists of the following steps. The input of this process is the pair of the syntactic and semantic structure of Japanese sentence and the partial syntactic structure of English translation equivalents.

1) **Identification of the pairs of Japanese word/phrase and their English equivalent word/phrase.**
I align the Japanese and English elements as described in Yamada et al. (1996) using the following information:

- bilingual dictionary for Japanese to English MT system, ALT-J/E
  This dictionary is used for the determination of pairs of equivalent word phrases of Japanese and English.
- English dictionary for English generation in ALT-J/E
  This dictionary is used when the suffix differs: for example the derivative, 'ing', between the English word in the bilingual dictionary entry and the English word in the English sentence within the aligned sentence pair.

[4]This step has not been implemented because anaphora resolution of English sentences needs detailed syntactic and semantic structure to achieve high accuracy. So, for now, I do not realize the anaphora resolution process in the whole system for the primary examination.

We remove function words such as prepositions, determiners and others from the target English sentence to find Japanese equivalent words/phrases in Japanese. This is because function words often must be changed depending on the types of head such as verb for preposition and noun for determiner in English.
As a result of this step, the alignment information of aligned words/phrases pairs are tagged for the Japanese and English structure . Furthermore, semantic information which was tagged in a Japanese word/phrase are also tagged for the English word/phrase which is aligned with the Japanese word/phrase.

2) **Identification of the candidates for antecedents of Japanese zero pronouns within the English sentence:**
In this step, the following English words/phrases are identified within the English sentence as possible translation equivalents:

- personal pronouns such as "I" or "you"
- "one".
- demonstratives such as "it" or "that"
- definite noun phrases such as a noun phrase with a definite article (e.g. "the company")

3) **Alignment of zero pronouns in Japanese sentences and translation equivalents of their referents in English sentences:**
The pairs of Japanese words/phrases and English equivalent words/phrases and the pairs of zero pronouns in the Japanese sentence and their antecedents in the English sentence are determined from: the candidates for the pairs of Japanese word/phrases and their English equivalent word/phrases which were identified at step 1; and the candidates for the antecedents of Japanese zero pronouns within the English sentence which were identified at step 2. This determination is conducted based on how strongly related the candidates are and how many pairs can be identified from them. The rules for the determination of the pairs of zero pronouns in the Japanese sentence and their antecedents in the English sentence are summarized as follows. The rules are extracted by hand based on the examination of the test set in Section 2.[5]

When applying the following rules for each zero pronoun, even if the rule is matched and the possible antecedent of the zero pronoun is determined, if the candidate does not satisfy the semantic constraints for the zero pronoun, the rule is not applied and the candidate is not determined as the antecedent of the zero pronoun. The rules are applied in the following order. If one rule is satisfied for each zero pronoun, the process of the following other rules application stops, except for rule 9.
For each sentence:

Rule 1 If a Japanese verb and an English verb are aligned
and if the subject equivalent[6] of the Japanese verb is an unaligned zero pronoun
and if the subject of the English verb does not have a Japanese equivalent,

[5]At the extraction process, I tried to make rules which can be identified fully automatically by using relatively simple Japanese and English syntactical information. I also take into account the coverage of each rules which can be covered as wide as possible.

[6]'the subject equivalent' means *ga* case (subject) or the case which will be translated as a subject in English by the MT system.

→ the subject of the English verb is determined to be the translation equivalent of the antecedent of the Japanese zero pronoun.

Rule 2 If a Japanese verb and an English verb are aligned
and if the subject equivalent of the Japanese verb is a zero pronoun
and if the zero pronoun and the subject of the English verb are aligned
and if the object equivalent[7] of the Japanese verb is an unaligned zero pronoun
and if the direct object of the English verb is identified as a antecedent candidate
and if the direct object of the English verb does not have a Japanese equivalent,
→ the direct object is determined as the translation equivalent of the antecedent of the Japanese zero pronoun.

Rule 3 If a Japanese verb and an English noun, derived from a verb, which is modified by a possessive pronoun, are aligned
and if a case of the Japanese verb is an unaligned zero pronoun
and if the Japanese equivalent of the possessive pronoun can not be determined,
→ the possessive pronoun is determined as the translation equivalent of the antecedent of the Japanese zero pronoun.

Rule 4 If the subject equivalent of a Japanese verb is an unaligned zero pronoun
and if the object equivalent of the Japanese verb is aligned with a case in the English sentence at subject position
and if the predicate part of the subject in the English sentence is in the passive voice,
→ there is no antecedent of the zero pronoun in the subject equivalent position within the English sentence because the passive voice expression is used and the zero pronoun is marked as unalignable.

Rule 5 If the subject equivalent of a Japanese verb is an unaligned zero pronoun
and the object equivalent of the Japanese verb is an unaligned zero pronoun
and if an English verb is in the passive voice
and if the subject of the English verb is identified as a candidate antecedent,
→ there is no antecedent of the zero pronoun in the subject equivalent position within the English sentence and the zero pronoun is marked as unalignable
and the antecedent of the zero pronoun in the object equivalent position is determined as the direct object of the English verb because the passive voice expression is used.

Rule 6 If the subject equivalent of a Japanese unit sentence in the Japanese sentence is an unaligned zero pronoun
and if the object equivalent of the Japanese unit sentence is an unaligned zero pronoun
and if the Japanese equivalents of a subject and a direct object within an English unit sentence in the English sentence can not be determined in the Japanese sentence or the Japanese unit sentence,
→ the translation equivalent of the antecedent of the zero pronoun in the subject equivalent position is determined as the subject of the English unit sentence
and the translation equivalent of antecedent of the zero pronoun in the object equivalent position is determined as the direct object of the English unit sentence.

Rule 7 If there is only one unaligned zero pronoun in the Japanese sentence
and if there is only one unaligned candidate antecedent in the English sentence,

[7] 'the object equivalent' means the *wo* case (direct object) or the case which will be translated as a direct object in English by the MT system.

→ the translation equivalent of the antecedent of the zero pronoun is determined as the candidate antecedent.
Or if there are more than one unaligned possible antecedents,
→ the translation equivalent of the antecedent of the zero pronoun is determined based on the following priority.

personal pronouns > "one" > demonstratives > definite noun phrases

Rule 8 If there is an unaligned zero pnonoun and all antecedent candidates have been aligned then, if there are one or more candidates which are aligned with Japanese elements in a different unit sentence to the unaligned zero pronoun,
→ the translation equivalent of the antecedent of the zero pronoun is determined from those candidates based on the priority in rule 7.

Rule 9 If the translation equivalent of an antecedent of a zero pronoun which is determined by rules 1 - 8, also has another Japanese equivalent within the Japanese sentence,
→ the zero pronoun is determined to be a zero pronoun with an intrasentential antecedent.

Rule 10 If there are any remaining unaligned zero pronouns,
→ these zero pronouns are ones whose antecedents are not explicitly translated in the English sentence and are marked as unalignable.

For example from the zero pronoun in the *ga* case (subject) in Japanese sentence in aligned sentence pair (1), rule 1 is applied and its antecedent is determined as the subject in the English sentence, "I".

4 Evaluation

4.1 Evaluation Method

The method to identify zero pronouns and their antecedents within Japanese and English aligned sentence pairs which was discussed in Section 3 was evaluated by automatically identifying zero pronouns and their antecedents from the functional test sentence set which is already aligned sentence by sentence. The conditions for the evaluation were as follows:

4.1.1 Evaluation of Target Sentence Pairs

The evaluation used Japanese and English aligned sentence pairs which contain zero pronouns (554 instances; 437 sentences) in a test set designed to evaluate the performance of Japanese-to-English machine translation systems (Ikehara et al., 1994) (3718 sentence pairs).

4.1.2 Analysis of Japanese and English sentence

For the sentence pairs which contain zero pronouns, the syntactic and semantic structure of each Japanese sentence was created using the Japanese-to-English MT system, ALT-J/E as the Japanese analyzer, described in section 3.1, and the partial syntactic structure of each English sentence was created using Brill's tagger as the English analyzer, described in section 3.2. To avoid any effects caused by problems at the analysis step on the evaluation of the rules to identify antecedents of zero pronouns, described in section 3.3, any incorrect structures in the Japanese and English sentences were modified by hand for the evaluation. Furthermore, to take into account the effect

of the anaphora resolution of anaphoric expression in English sentences on the accuracy of the identification of the antecedents of zero pronouns, the antecedents of anaphoric expression in the English sentences were determined by hand and the accuracy with and without anaphora resolution was compared.

4.1.3 Antecedents

For the sentence pairs which contain zero pronouns, antecedents of each zero pronoun within the text set were automatically identified (Section 3).

4.1.4 Evaluation Parameters

To examine the effectiveness of automatically identifying antecedents of Japanese zero pronouns within English translations, the accuracy of the identified antecedents of each zero pronoun of the following three types were examined.

- Zero pronouns with intrasentential antecedents (150 instances)
- Zero pronouns with deictic referents which are not expressed in English translation by using passive voice expressions (157 instances)
- Zero pronouns with deictic referents which are expressed in English translation (247 instances)

4.1.5 Successfully Identified Antecedents of Zero Pronouns

When rules in Section 3.3 correctly identify the accurate antecedent of a Japanese zero pronoun, with intrasentential antecedents or with deictic referent, which are expressed in the English translation, or when rules in Section 3.3 correctly identify a Japanese zero pronoun as unalignable, that is there is no deictic referents expressed in the English translation, the antecedents of zero pronoun are judged to be identified successfully.

4.2 Identification Accuracy

The accuracy of identified antecedents of zero pronouns is shown in Table 2. As shown in this table, the accuracy of identified antecedents for three types of zero pronouns using rules described in Section 3.3 is as high as 91.5% in the test using an English analyzer with anaphora resolution and 87.2% even in the test using an English analyzer without anaphora resolution. In particular, the identification of unalignable for zero pronouns with deictic referents which are not expressed in English translation is 100% accurate in both tests. Furthermore, according to these results, the anaphora resolution in English only affects the accuracy of antecedent identification for the zero pronouns with intrasentential antecedents (89.3% with anaphora resolution, 73.3% without anaphora resolution). This result shows that even without using anaphora resolution at English analysis, this method achieves relatively high accuracy for zero pronouns with intrasentential antecedents.

The result of the detailed examination of the zero pronouns whose antecedents can not be identified correctly using this method (47 instances) are summarized in Table 3. The most common cause of error is that the Japanese sentence was translated freely and there is no corresponding antecedent within the English translation (42 out of 47 instances). For these sentence pairs, the antecedents of the zero pronouns are not explicitly expressed in the human translation, but the Japanese analyzer still needs the resolution of the zero pronouns because the machine translation system which contains the Japanese analyzer can not translate freely like the human translator.

Table 2: Identification accuracy of antecedents for types of zero pronouns

condition at English analysis	Intrasentential	Deictic		Total
		passive	explicit	
with anaphora resolution	89.3% (134/150)	100.0% (157/157)	87.4% (216/247)	91.5% (507/554)
without anaphora resolution	73.3% (110/150)	100.0% (157/157)	87.4% (216/247)	87.2% (483/554)

So, this problem is caused by the methodological limitation of proposed method. The remaining 5 instances are caused by zero pronouns with intrasentential antecedents. In 4 out of 5 cases the antecedents of the zero pronouns were not explicitly translated in English by using the passive voice but humans can easily understand that the antecedents of the zero pronouns are in the same sentence. This problem is also caused by the limitations of proposed method. The remaining case is caused by the fact that the verb of the zero pronoun and the verb of their antecedents could not be aligned. This is not a limitation of proposed method but a limitation of the alignment algorithm. These results shows that, for the test set, 508 out of 554 zero pronouns can have their antecedents identified using proposed method methodologically, and almost all of the antecedents of zero pronouns which can be identified from the Japanese and English aligned sentence pairs are correctly identified by using proposed method (99.8% ; 507 out of 508 instances).

Table 3: The Cause of the Error in Automatic Identification of antecedents

the cause of the error	Intrasentential	Deictic(explicit)	Total
free translation	11	31	42
passive voice	4	0	4
predicate can not be aligned	1	0	1

According to these results, the proposed method is effective for the automatic identification of Japanese zero pronouns and their antecedents from Japanese and English aligned sentence pairs and by using a large amount of aligned sentence pairs it is possible to identify antecedents of Japanese zero pronouns for almost all types of Japanese zero pronouns.

5 Conclusion

This paper proposes a powerful method for the automatic identification of Japanese zero pronouns and their antecedents from Japanese and English aligned sentence pairs. In this paper, only Japanese and English language pairs have been discussed. But this method can be applied to various kinds of language pairs such as Italian and English, and the effectiveness of the identification depends on how different the two languages are. In the future, the effectiveness of proposed method for aligned sentence pairs with zero pronouns with intersentential antecedents and for very large corpora of aligned sentence pairs will be examined. The effectiveness of the use of English parsers for the English analysis will also be considered. Methods for extracting the most effective rules for resolving Japanese zero pronouns from aligned sentence pairs will also be studied. Furthermore, I would like to link more powerful and more robust word alignment techniques, which can align for unknown words in noisy texts, to the proposed method and also would like to link the proposed method with sentence alignment techniques.

Acknowledgments

I would like to thank Jun'ichi Tsujii for helpful discussion of many of the ideas and proposals presented here during my stay at UMIST from September 1995 to September 1996. I am also grateful to Francis Bond and several anonymous reviewers of WVLC-5 for helpful comments on earlier drafts of the paper.

References

Eric Brill. 1992. A simple rule-based part of speech tagger. In *Proc. of ANLP92*, pages 152–155, ACL.

Peter F. Brown, Jennifer C. Lai and Robert L. Mercer. 1991. Aligning sentences in parallel corpora. In *Proc. of 29th Annual Meeting of ACL*, pages 169–176, ACL.

Kenneth W. Church. 1993. Char-align: A program for aligning parallel texts at the character level. In *Proc. of 31st Annual Meeting of ACL*, pages 1–8, ACL.

Ido Dagan, Alon Itai and Ulrike Schwall. 1991. Two languages are more informative than one. In *Proc. of 29th Annual Meeting of ACL*, pages 130–137, ACL.

Ido Dagan and Kenneth W Church. 1994. Termight: Identifying and translating technical terminology. In *Proc. of ANLP94*, pages 34–40, ACL.

Kouji Dousaka. 1994. Identifying the Referents if Japanese Zero-Pronouns based on Pragmatic Condition Interpretation. In *Trans. of IPS Japan*, 35(10):768–778. In Japanese.

Pascale Fung and Kenneth W. Church. 1994. K-vec: A new approach for aligning parallel texts. In *Proc. of COLING94*, pages 1096–1102.

William A. Gale and Kenneth W. Church. 1991. A program for aligning sentence in bilingual corpora. In *Proc. of 29th Annual Meeting of ACL*, pages 177–184, ACL.

Satoru Ikehara, Shirai Satoshi and Kentaro Ogura. 1994. Criteria for Evaluating the Linguistic Quality of Japanese-to-English Machine Translation. In *Journal of JSAI*, 9(5):569–579.

Satoru Ikehara, Shirai Satoshi, Akio Yokoo and Hiromi Nakaiwa. 1991. Toward MT system without Pre-Editing -Effects of New Methods in ALT-J/E-. In *Proc. of MT Summit III*, pages 101–106.

Masahiko Haruno and Takefumi Yamazaki. 1996. High-Performance Bilingual Text Alignment using Statistical and Dictionary Information In *Proc. of 34th Annual Meeting of ACL*, pages 131–138.

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

Martin Kay and Martin Roscheisen. 1993. Text-translation alignment. In *Computational Linguistics*, Vol. 19, No. 1, pages 121–142.

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

Julian Kupiec. 1993. An algorithm for finding noun phrase correspondences in bilingual corpora. In *Proc. of 31st Annual Meeting of ACL*, pages 17–22, ACL.

Yuji Matsumoto, Hiroyuki Ishimoto and Takehito Utsuro. 1993. Structural matching of parallel texts. In *Proc. of 31st Annual Meeting of ACL*, pages 23–30, ACL.

Masaaki Murata and Makoto Nagao. 1997. An Estimation of Referents of Pronouns in Japanese Sentence using Examples and Surface Expressions. In *Journal of Natural Language Processing*, 4(1):87–109, Association of Natural Language Processing. In Japanese.

Hiromi Nakaiwa. 1997. 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* To appear.

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 ANLP92*, pages 201–208, ACL.

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 COLING94*, pages 672–678.

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 TMI95*, pages 96–105.

Hiromi Nakaiwa and Satoru Ikehara. 1996. Anaphora Resolution of Japanese Zero Pronouns with Deictic Reference. In *Proc. of COLING96*, pages 812–817.

Tetsuya Nasukawa. 1996. Full-text processing: improving a practical NLP system based on surface information within the context. In *Proc. of COLING96*, pages 824–829.

Kentaro Ogura, Akio Yokoo, Satoshi Shirai and Satoru Ikehara. 1993. Japanese to English machine translation and dictionaries. In *Proc. of 44th Congress of the International Astronautical Federation*, paper No. IAA.5.1-93-720.

Hideki Tanaka. 1994. Verbal case frame acquisition from a bilingual corpus: Gradual knowledge acquisition. In *Proc. of COLING94*, pages 727–731.

Dekai Wu. 1995. An algorithm for simultaneously bracketing parallel texts. In *Proc. of 33rd Annual Meeting of ACL*, pages 244–251, ACL.

Marilyn Walker, Masayo Iida and Sharon Cote. 1990. Centering in Japanese discourse. In *Proc. of COLING90.*

Setsuo Yamada, Hiromi Nakaiwa, Kentaro Ogura and Satoru Ikehara. 1995. A method of automatically adapting a MT system to different domains. In *Proc. of TMI95*, pages 303–310.

Setsuo Yamada, Hiromi Nakaiwa and Satoru Ikehara. 1996. A new method of automatically aligning expressions within aligned sentence pairs. In *Proc. of NeMLaP2*, pages 56–65.

Kei Yoshimoto. 1988. Identifying zero pronouns in Japanese dialogue. In *Proc. of COLING88*, pages 779–784.