Automatic Detection of Antecedents of Japanese Zero Pronouns Using a Japanese-English Bilingual Corpus

Dong Zhan 1301211064@pku.edu.cn
School of Software and Microelectronics, Peking University, Beijing, 100871, China

Hiromi Nakaiwa nakaiwa@is.nagoya-u.ac.jp
Graduate School of Information Science, Nagoya University, Aichi, 464-8601, Japan

Abstract

In this paper we present a method of detecting zero pronouns in Japanese clauses and identifying their antecedents using aligned sentence pairs from a Japanese-English bilingual corpus and open resource tools. We use syntactic and semantic structures and the alignment of words and phrases in the sentence pairs to automatically detect zero pronouns and determine their antecedents using English translations. We build rules to link antecedents with zero pronouns and create filters to remove problematic sentence pairs. Experimental results confirm the effectiveness of our method. The proposed method allows the construction of an annotated corpus of zero pronoun sentences in which the antecedents of the missing pronouns are flagged. This would be very useful for machine translation (MT), because zero pronoun detection is a vital problem when translating languages which allow zero pronouns.

1 Introduction

In many languages, such as Japanese and Chinese, elements which can be easily deduced by the reader are frequently omitted from subsequent expressions in discourses. Omissions related to obligatory subject and object cases are often referred to as zero pronouns. A zero pronoun can be thought of as a noun phrase which is obligatory but which is not expressed explicitly. Zero pronoun resolution is more complicated than overt pronoun resolution because a natural language processing (NLP) system has to detect zero pronouns before it can identify their antecedents. Determination of the antecedents of zero pronouns is vital in many NLP applications, including information extraction, question answering, machine translation, etc.

Many algorithms have been proposed for pronominal anaphora resolution [9, 2, 10]. However, resolving the anaphora problem is much harder in languages which allow zero pronouns compared to languages which do not allow zero pronouns, such as English, since detection of zero pronouns should be done before anaphora resolution. Several methods have been proposed to solve this problem [12, 3, 14]. Some of these methods use semantic and pragmatic constraints, such as semantic constraints on cases, modal expressions, or verbal semantic attributes, to determine the referents of zero pronouns [12]. Other methods use a machine learning approach for anaphora resolution, such as using semantic attributes as computable features to perform identification and resolution

of anaphoric zero pronouns in Chinese [14] or using syntactic patterns of zero pronouns and their antecedents as features for zero-anaphora resolution in Japanese [3, 5].

Most of these methods focus on zero pronoun resolution and detection of antecedents of zero pronouns as a crucial part of zero pronoun resolution. All of these methods employ either an annotated corpus containing zero pronoun sentences, with anaphoric relationships tagged by annotators [3, 4], or an annotated corpus provided by other persons or institutes [14, 1]. There are many ways to construct such an annotated corpus. Most of these corpora are monolingual and are annotated by hand. Manual annotation of these corpora can be very time consuming and may require relevant knowledge, so in this study we attempt to find a way to automatically detect the antecedents of zero pronouns.

Different languages contain different types of anaphoric expressions. Since similar languages often contain similar anaphoric expressions, languages which are very different from each other, such as English and Japanese, generally differ more markedly in regard to how they use pronouns and how these pronouns are linked to their antecedents. Thus, if we use a bilingual corpus, we are able to contrast how each language handles corresponding occurrences of subjects, objects and pronouns in the same sentence, and we can utilize the translation to identify the antecedents of zero pronouns.

Some research on automated corpus annotation using a bilingual corpus has been conducted [11]. The purpose of that study was to formulate rules for the anaphoric resolution of Japanese zero pronoun sentences using aligned sentence pairs. The same method was also used in another study, in which a valency transfer dictionary which contained 16,000 pairs of Japanese case-frame patterns was used for zero pronoun detection [6]. It showed the relationship between corresponding Japanese verbs and their English translations using different semantic models. For example, ”yomu” in the patterns ”N1(human)-ga N2(abstract thing)-o yomu” should translate as ”N1 read N2” in English, but in the pattern ”N1(subjects)-ga N2(”vote”)-o yomu” should translate as ”N1 predicts (the outcome of the vote will be) N2”. This dictionary makes it easy to find the obligatory elements of Japanese sentences and it can also be used to automatically recognize intransitive verbs. But since there were no open source parsers for this dictionary, this method of zero pronoun resolution could not be widely implemented. Several years have passed since this method was first proposed however, and in that time many useful tools for natural language processing have been created which have achieved good accuracy in tasks such as morphological analysis, parsing and alignment. These tools are widely used in the field of NLP and most of them are open source software. Therefore, in this study we propose a method of automatically annotating Japanese zero pronouns and their antecedents into a Japanese-English bilingual corpus, using open source tools.

2 Task Definition

For the purpose of machine translation, it is important to recognize that some grammatical elements not present in one language could be obligatory elements in another language. In particular, subjects and objects are often omitted in Japanese, but are obligatory in English. So it is natural to consider using the differences between Japanese and English to resolve the issue of missing pronouns in Japanese. A Japanese-English bilingual corpus could be useful if the antecedents of zero pronouns are difficult to recognize in Japanese sentences but these missing pronouns have equivalent antecedents in English sentences. For example, subject is zero pronoun in Japanese sentence of example (1) and it has a equivalent antecedent ”I” in English sentence, then zero pronoun

can be aligned with equivalent antecedent in English.
(1) 私は疲れているとき、本を読みたい。

Literally translated, "When I am tired, want to read book." In grammatically correct English, it would be translated as, "When **I** am tired, (I) want to read a book." The first, bold "I" is the antecedent, and in this sentence it allows us fill in the "I" in parenthesis, which is the English equivalent of the missing zero pronoun in Japanese.

So in this study, we propose a method to detect the antecedents of Japanese zero pronouns using English translations of Japanese sentences. This method involves two steps. First, we need to detect whether or not the original sentence contains zero pronouns, and if it has zero pronouns we want to know where the zero pronouns occur. Then we try to detect the antecedents in the English translation and decide which antecedent corresponds to the missing zero pronoun. The translation process may result in some errors and unsuitable sentences, however, so we have to detect and remove unsuitable sentence pairs. Zero pronouns can be divided into three types, based on the location of the antecedent:
The antecedent of the zero pronoun is located in sentence futher back, mostly in the previous sentence:
(2) 彼はとてもいい人です。昨日、(ϕ-ga) 私に手伝ってくれました。
He is a good person. Yesterday helped me.
The antecedent of the zero pronoun is located in the same sentence:
(3) 彼は宿題を終わって、(ϕ-ga) テレビを見ました。
He finished homework, watched TV.
The antecedent of the zero pronoun does not exist at all in context:
(4)(ϕ-ga) 東京にいきたい。　　Want to go to Tokyo.

In this paper we focus on both simple sentences and complex sentences, which contain clauses in addition to one simple sentence. But to detect the antecedent of a zero pronoun within a clause of a complex sentence or in a previous sentence, we will need to combine our translation comparison method with other anaphora resolution methods. For example, our method can identify that the antecedent of the zero pronoun in the second clause of sentence (2) is "he", but it does not know that "he" refers to the character "彼" in the previous clause. If we can find the antecedent of a zero pronoun in an English translation, we can also determine its referent using anaphora resolution on the translation, and then align the referent and the corresponding Japanese word. This will allow us to resolve even intra-sentential zero pronoun anaphora. There have been many studies on anaphora resolution in English, but in this study we only focus on detecting the antecedents of zero pronouns in English translations of Japanese sentences.

3 Corpus

For this study, we used the Japanese-English Scientific Paper Excerpt Corpus (ASPEC-JE)[1], which contains 3,008,500 parallel sentence translations collected from Japanese-English scientific paper abstracts. There are no syntactic or semantic structures associated with the sentences in this corpus, so we need to do segmentation, Part-Of-Speech tagging and parsing. We removed especially long sentences from the corpus (more than 150 characters in one sentence) because the syntactic structures of these sentences were too complex to be used. We removed 486,958 long sentences, and then used the remaining 2,521,542 sentence pairs to train the Japanese to English alignment model. To examine the effectiveness of this method, only 1,000 sentences were used for a closed test and 200 sentences for an open test, because this seems to be a reasonable number of

[1] http://lotus.kuee.kyoto-u.ac.jp/ASPEC/

sentences which could be evaluated manually. All 1,200 of these sentences were chosen randomly.

4 Linking zero pronouns and their antecedents by aligning Japanese-English sentence pairs

Our method can be divided into several steps: 1) morphological analysis of the Japanese-English corpus (MeCab[2][8] for Japanese and Stanford segmentation[3] for English); 2) alignment of Japanese words and English words (GIZA++[4][13]); 3) syntactic and semantic analysis of Japanese sentences; 4) syntactic analysis of English sentences; 5) identification of unsuitable sentence pairs; and 6) linking of Japanese zero pronouns with their antecedents. For Japanese syntactic and semantic analysis, we use two tools, CaboCha[5][7] for syntactic analysis and SynCha[6][5] for Japanese predicate-argument structure analysis. These tools allow us to recognize which part of a sentence is the subject, object or predicate and to identify the dependency relationships between word phrases. For English syntactic analysis, the Stanford parser[7] is used. There are two types of output; context-free, phrase structure grammar representations and universal dependencies. We use the latter to build syntactic relationship structures. Overview of the process is shown in Figure 1.

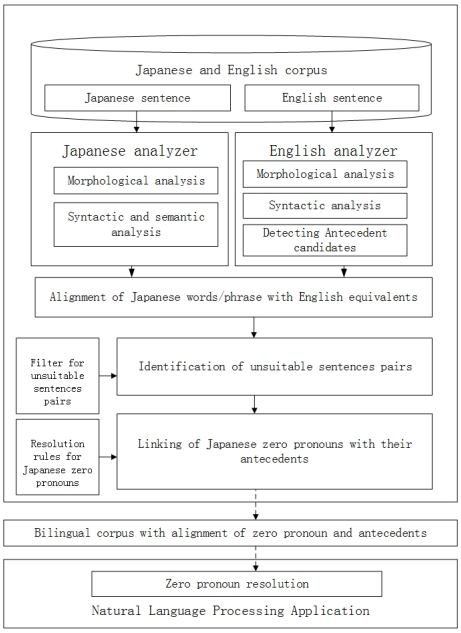

Figure 1: Overview of the proposed method

4.1 Identification of antecedent candidates in English sentences

Before linking Japanese zero pronouns with their antecedents as indicated in English translations, we need to specify a range of antecedent candidates. Anaphoric expressions are common in English, and there is a set of special identity words which tend to often appear as antecedents. In order to not miss these common antecedents, we must consider not only definite nouns and noun phrases, such as "the company", but also these identifying words, as possible translation equivalents. These identifying words include personal pronouns (I, you, he, she, they, we, it); impersonal words (one, everyone, no one, men, women, human beings, people) and demonstratives (this, that, these, those, each, every).

[2]http://taku910.github.io/mecab/

[3]http://nlp.stanford.edu/software/tokenizer.shtml

[4]http://www.statmt.org/moses/giza/GIZA++.html. We use alignments of translations in both directions (from Japanese sentences to English sentences and vice versa) and combine them for the final alignment. This means Japanese words or phrases and their English counterparts are only linked if they are aligned in the translation results in both directions.

[5]http://taku910.github.io/cabocha/

[6]http://www.cl.cs.titech.ac.jp/ ryu-i/syncha/

[7]http://nlp.stanford.edu/software/lex-parser.shtml

4.2 Identification of unsuitable sentence pairs

There are situations where the aligned sentence pairs are not suitable for annotating antecedents of zero pronoun. These sentences need to be identified and moved out. For example, when English parser was unable to make a full syntactic structure, we cannot use it for identification of antecedent candidates. Considering the accuracy rate of tools we used and the quality of the bilingual corpus, aligned sentence pairs are identified as unsuitable if they have the problems caused by corpus, such as freely translation, or errors caused by tools, like parser and alignment errors. For identifying these unsuitable sentences, we have made some rules to filter these sentences out automatically after having analyzed the result of original rules. The details of these filters will be explained in section 5.3.

4.3 Zero pronoun detection

Zero pronoun sentences occur frequently in Japanese and zero pronouns usually appear where either the grammatical subject or object would appear. Moreover, understanding the referent of the zero pronouns in these positions is vital for sentence comprehension. Hence, we focus on zero pronouns which appear as subjects and objects. Normally, a subject is an essential component of a complete sentence, unless the sentence is formulated in a passive voice ("The mailman was attacked.") or in the imperative mood ("Attack the mailman!"). If a Japanese sentence does not have a subject, and the sentence is not in the passive voice or imperative mood, we assume there is a zero pronoun in subject position. Regarding objects, our assumption is based on whether the predicate contains a transitive or intransitive verb. If the predicate of a Japanese sentence contains a transitive verb but there is no object, then it is assumed that there must be a zero pronoun in object position. These assumptions are the basis of zero pronoun detection in Japanese sentences. However, we are using an open source parser which cannot recognize intransitive verbs in Japanese sentences. Furthermore, the same verbs tend to be transitive or intransitive in English as in Japanese. But with the help of an English translation, we can recognize intransitive verbs in Japanese sentences by noting if the corresponding English verb in the sentence has no direct object.

4.4 Linking Japanese zero pronouns with their antecedents

Now that we have ways to filter out unsuitable sentences and to identify zero pronouns, we need a way to link an antecedent which appears in an English sentence to a zero pronoun in a Japanese sentence. Nakaiwa proposed ten rules for this purpose (1999), and we used these rules as our starting point. However, based on an analysis of the results and the features of the corpus, we decided to formulate new rules to improve performance. The ten rules proposed in Nakaiwa's paper can be described as follows:

Rule 1: Alignment between subjects
Rule 2: Alignment between objects
Rule 3: Alignment of Japanese subjects and English possessive pronouns
Rule 4: Unalignment of Japanese subject when the passive voice is used in English
Rule 5: Non-alignment of Japanese subjects, and alignment of Japanese objects with English subjects, when a passive voice is used in English
Rule 6: Alignment of both subjects and objects
Rule 7: First default alignment rule for remaining unaligned zero pronoun
Rule 8: Second default alignment rule for remaining unaligned zero pronouns
Rule 9: Default unalignment rule for unaligned zero pronouns
Rule 10: Identification of intra-sentential antecedents of any aligned zero pronouns

These rules are applied from 1 to 10 in numeric order. If a rule is satisfied for a zero pronoun, the process will stop. Each rule has its own application conditions, and a rule is applied only if all of the conditions for the rule are satisfied. Details of Rule 4, Rule 5, Rule 7 and Rule 9 are shown below, since the results of these rules will be analyzed in the following section:

Rule 4: Non-alignment of Japanese subjects when a passive voice is used in English
IF Sj of Pj is an unaligned ϕ & Oj of Pj is aligned with Se of USe & Pe of USe is in a passive voice
THEN there are no antecedents of Sj because a passive voice is used in the English translation: unalignable

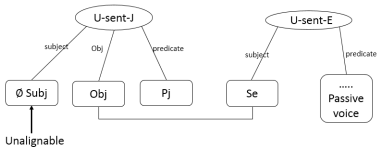

Figure 2: Rule4

For example, in sentence pair (5), the process for Rule 4 is as follows:
(5) (ϕ-**ga**) 内部リレーの使い方を説明した。
The usage of the inside relay was explained .
Oj: 使い方 <-> Se: usage & Pe: "was explained" is a passive voice => Sj: ϕ -ga = unalignable

Rule 5: Non-alignment of Japanese subject, and alignment of Japanese object with English subject, when a passive voice is used in English
IF Sj of Pj is an unaligned ϕ & Oj of Pj is an unaligned ϕ & Pe of USe is in a passive voice & Se of USe is an antecedent candidate Ai
THEN there are no antecedents of Sj because a passive voice is used in the English translation: unalignable & align ϕ-Oj with Se

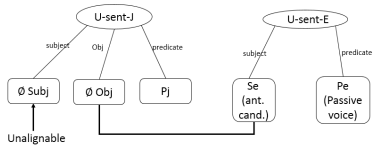

Figure 3: Rule5

(6) (ϕ-**ga**)(ϕ-**o**) 外用剤で治療した。
It was treated by the external preparation .
Pe: "was treated" is a passive voice & Se: it is an antecedent candidate Ai => Sj: ϕ-ga = unalignable & Oj: ϕ-o <-> Se: it

Rule 7: Default alignment rule for remaining unaligned zero pronouns
IF there is only one unaligned zero pronoun case in Japanese and there are one or more unaligned candidate antecedents
THEN we determine the antecedent of zero pronoun based on the following priority: personal pronoun > one > demonstratives > definite NP

Rule 9: Default non-alignment rule for unaligned zero pronouns
IF there are any remaining cases of unaligned zero pronouns in Japanese
THEN these are determined to be ϕ-Cj's whose antecedents are not explicitly translated in USe and are marked as unresolved

5 Experiment

5.1 Evaluation

We evaluated the proposed method using aligned sentence pairs from the ASPEC-JE corpus. A closed test was conducted to evaluate the performance of the rules proposed by [11]. After analyzing the results, we added several filters to remove unsuitable sentence pairs with errors caused by processing tools or which had inaccurate translations. We then added additional rules to detect the antecedents of zero pronouns. A closed test

was again performed on the same data sets, and the process was repeated several times. Finally, an open test was conducted on unseen data to evaluate the effectiveness of the proposed method with an unknown corpus.

5.2 Closed test of Nakaiwa method

We first evaluated the effectiveness of Nakaiwa's original 10 rules for detecting of zero pronoun antecedents. Before the evaluation, sentences without zero pronouns should be removed, and according to the parsing results, 95 sentences without zero pronouns were removed automatically, since these types of sentences are not our target. The remaining sentences were then examined using Nakaiwa's original ten rules. Some rules, such as Rule 5, Rule 6 and Rule 9, determine the alignment or non-alignment of zero pronouns more than once at the same time. Rule 5 and Rule 6 process zero pronouns in both the subject and object positions of the same sentence, for example. Rule 9 identifies sentences which are unalignable. In order to evaluate our results, we had to check the output manually. The results are shown in Table 1.

Rule	Times rule applied correctly/Times rule applied (Accuracy)	Number of zero pronouns determined and linked correctly/Number of zero pronouns determined (Accuracy)
1	95/96 (99.0%)	95/96 (99.0%)
2	10/11 (90.9%)	10/11 (90.9%)
3	0/0 (-)	0/0 (-)
4	333/333 (100%)	333/333 (100%)
5*	17/104 (16.3%)	121/208 (58.2%)
6*	1/2 (50%)	2/4 (50%)
7	28/59 (47.5%)	28/59 (47.5%)
8	9/12 (75.0%)	9/12 (75.0%)
9	85/424 (20%)	85/502 (16.9%)
10	0/0 (-)	0/0 (-)
Totals	578/1041 (55.5%)	683/1225 (55.8%)

Table 1: Results using Nakaiwa's original rules

* Correct results means both subject and object zero pronouns were correctly determined.

5.3 Analysis of results

From these results we can see that only 55.5% of applied rules determined zero pronouns and their antecedents correctly, and that the overall accuracy rate of zero pronoun and its antecedent determination was 55.8%. Rule 1, for determining zero pronouns in the subject position, achieved a precision rate of 99.0%,with the one incorrect result caused by a Japanese parser error. Rule 2, for determination of zero pronouns in the object position, achieved 90.9% precision. Rule 4, for detection of passive voices, achieved 100% precision. These results show that the original method is successful for many sentences. However, there were no or very limited cases in which Rule 3, Rule 6 and Rule 10 could be applied.

Rule 5, for both detection of zero pronouns in the object position and detection of passive voices, achieved only 58.2% precision for zero pronoun determination and Rule 9 achieved very poor precision (16.9%) for zero pronoun determination, which differs markedly from Nakaiwa's own results (1999), in which accuracy rates of 92.9% and 44.8% for Rule 5 and Rule 9 were achieved, respectively. After checking the results for Rule 5, we found that in most cases the English translations were in a passive voice. These cases should have been processed by Rule 4, but went unrecognized because there are no objects of Japanese sentences which do not satisfy the condition of Rule 4. As

for Rule 9, its poor performance was due to its inability to detect intransitive verbs. Thus, we clearly need to adjust the rules in order to improve accuracy.

We also want to determine the recall of our method. To do this, we need to know the actual number of zero pronouns in the sentences used in the closed test. As mentioned in section 4.3, we can determine whether zero pronouns exist or not based on whether subjects or objects exist in the sentences. The number of zero pronouns in the subject position is counted automatically, but since our system cannot determine whether a Japanese verb is transitive or intransitive automatically, we need to check the number of zero pronouns in the object position manually. Based on whether or not the subject or object is missing, all of the sentences in the closed test can be divided into one of four types:

Type of sentence (automatically determined)	Number of simple sentences	Zero pronouns in subject position	Number of zero pronouns in object position
No zero pronoun	95	0	0
Subject is missing, object exists	490	490	0
Subject is missing, object unknown	231	231	33
Subject exists, object unknown	278	0	8

Table 2: Distribution of zero pronouns

The total number of zero pronouns in the closed test corpus can be calculated as follows: 490+231+33+8 = 762. We can then calculate the recall of the result: 676/762 = 88.7%. Since there were many incorrect determinations of zero pronouns in our experiment, we tried to analyze these errors to improve accuracy. In our analysis of incorrect determinations of default results for Rule 5 (87 cases), Rule 7 (31 cases) and Rule 9 (339 cases), we found the following types of errors:

(A) There is no explicit antecedent in the English sentence due to faulty translation.
(B) Japanese parser error. Syntactic and semantic analysis may have caused errors in detecting semantic relationships.
(C) English parser error, such as syntactic and semantic analysis errors.
(D) Alignment error. Words that should be aligned cannot be found or were mismatched.
(E) No antecedent candidates.
(F) Cases where Japanese verbs were translated into continuous tenses or as gerunds in English, leading to incorrect predicate classification.
(G) Incorrectly detected zero pronouns in the object position.
(H) Intransitive verbs which act as predicates in Japanese sentences. In such cases, there is no object in the sentence but the system cannot detect this automatically. This is caused by the absence of an open source tool to determine if a Japanese verb is transitive or intransitive.

	Total	A	B	C	D	E	F	G	H
Rule5	87	0	0	0	0	0	0	87	0
Rule7	31	15	3	4	0	0	9	0	0
Rule9	339	7	11	4	8	1	6	0	302

Table 3: Distribution of errors by type of error for Rule 5, Rule 7 and Rule 9

The statistical results are shown below in Table 3. According to the results, all of the Rule 5 errors were caused by incorrect detection of zero pronouns in the object position, and most of the Rule 9 errors were caused by faulty intransitive verb detection. Poor translations (type A) and parser errors (Type B and Type C) were responsible

for many of the errors. By analyzing these errors, we can adjust the rules to improve performance.

5.4 Modification of original rules

We found that all of the Rule 5 errors were caused by false detection of zero pronouns in the object position which did not exist, and that most of these sentences were translated into a passive voice. Such sentences should have been processed using Rule 4, but since these sentences do not have objects, they do not satisfy the conditions of Rule 4. This is the result of a syntactic difference between Japanese and English.
(9) 標記課題について検討した。 The problem with the title was examined .

For example, in sentence pair (9), the corresponding English translation is in a passive voice, and the noun phrase before “について” should be the object of the English sentence (“the problem with the title”). In Japanese “について” often connects noun phrases and predicates using the pattern “noun + について + verb” . The noun acts as an object in this case, but the Japanese parser cannot recognize it. So we developed a new rule to handle these situations, hoping this would improve our results. This new rule is not for the detection of antecedents of zero pronouns, but is instead a complement of tools to recognize whether there is a zero pronoun in the object position.

Rule11: Detection of zero pronouns as objects
IF Oj of Sj are unaligned ϕ & Cj in Noun+ について modifies the predicate
THEN replace ϕ-Oj with Cj

(10) 標記課題について検討した。 The problem with the title was examined .
Oj: ϕ-o & Nj (標記課題) について exists & Nj (標記課題) modifies 検討した。=> Replace ϕ-o with Nj (標記課題)

Rule	Times rule applied correctly/Times rule applied (Accuracy)	Number of zero pronouns determined and linked correctly /Number of zero pronouns determined (Accuracy)
1	95/96 (99.0%)	95/96 (99.0%)
2	10/11 (90.9%)	10/11 (90.9%)
3	0/0 (-)	0/0 (-)
4	471/471 (100%)	471/471 (100%)
5	17/17 (100%)	34/34 (100%)
6	0/1 (0%)	0/1 (0%)
7	35/64 (54.7%)	35/64 (54.7%)
8	4/5 (80.0%)	4/5 (80.0%)
9	61/360 (16.8%)	61/371 (16.4%)
10	0/0 (-)	0/0 (-)
Totals	693/1025 (67.6%)	710/1053 (67.4%)

Table 4: Result of method with added rule

We insert this new rule bewteen Rule 2 and Rule 3 (not showed in Table 4). We repeated the experiment. The results of this experiment are shown in Table 4. We can see that the total number of sentences decreased from 1,041 to 1,025. This is because in the 16 sentences in which a subject existed but an object did not exist, the object was added by applying Rule 11, so that no zero pronouns remain in these sentences. Thus, they will not be processed by the other rules and will not be counted in the total number of detected zero pronoun sentences. A total of 173 sentences were processed by Rule 11. The number of sentences processed by Rule 4 increased from 333 to 471, and the largest increase in processing occurred with Rule 5 (87 sentences), even though there are sentences which Rule 5 cannot be applied to, because they do not satisfy the

condition of having antecedents in English, in which case the default rules are applied (51 cases). These sentences are now processed by Rule 4 after Rule 11 was added. As a result, sentences being processed by Rule 5 and by the default rules have decreased. We can also see that the total precision of zero pronoun determination increased from 61.6% to 67.4% and that recall increased from 88.7% to 93.2% (710/762).

5.5 Importation of filters

From our analysis of Table 3, we could see that many errors originated before processing, such as errors caused by incorrect translations. It is hard to avoid these types of errors, but we can add filters to detect problematic sentences and remove them in advance.

In some cases, part of the original Japanese sentence is translated into a noun phrase or gerund. Sentence pairs which had different numbers of unit sentences were rejected as unsuitable in Nakaiwa's research (1999). We use a similar rule, but we only filter out unsuitable sentence pairs when English translation contains fewer unit sentences than the Japanese original so that if the English translation contains more unit sentences than the Japanese original, we still try to find the antecedent of a zero pronoun by trying to align the zero pronoun with possible antecedents in each of the unit sentences.

Filter 1: Filter for sentence pairs when English translation contains fewer unit sentences than the Japanese original. If English translation contains fewer unit sentences than the Japanese original, then mark the sentence pair as unsuitable.

(11) 柔軟性を重視してＪＡＶＡを用いた。JAVA was used because of its flexibility.

Pj: 2 unit sentences & Pe: 1 unit sentence => unsuitable

A large number of errors were caused by faulty detection of intransitive verbs, described as Type H errors in Table 3, thus we built a filter to determine if there is a zero pronoun in the object position.

Filter 2: Filter for zero pronoun in the object position with intransitive verb
If there is no object in the English sentence, and a subject and predicate exist in the Japanese sentence which are aligned with the English predicate, and no unaligned words remain, then there is no zero pronoun in object position.

(12) 画素構造の改造が（ϕ-o）進んでいる。Modification of the pixel structure is progressing .

Pj: 進んでいる <-> Pe: has advanced .& Sj: 改造が exist & object not exist in English sentence => Sj: Obj is not zero pronoun

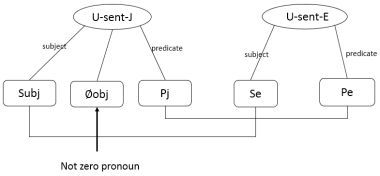

Figure 4: Filter2

Filter 3 is used to detect errors caused by the limitations of Japanese language analysis tools. For example, in sentence(13), the object “事実”(”fact”) does exist in the Japanese sentence, but the Japanese parser cannot determine that the word before “は” is an object.

Filter 3: Filter for non-alignment of zero pronouns in the object position
If subject, object and predicate all exist in the English sentence and are aligned with Japanese words, and no words are remain to serve as antecedents of the object zero pronoun, then there is no zero pronoun in object position.

(13) この事実は著者が既にこのシリーズで (ϕ-ga) 論じてきた。 The author has already discussed this fact in this series . Pj: 論じてきた <-> Pe: has already discussed & Sj: 著者が <-> Se: The author & Oe: fact <-> other words in Japanese & no object left in English => unsuitable sentence pair &Sj: Obj is not a zero pronoun

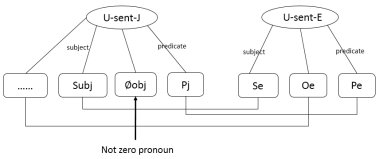

Figure 5: Filter3

Filter 4: Detecting mismatched sentence pairs. In a case where both the Japanese and English sentences contain sub-sentences, if the predicate and subject of the Japanese sub-sentences are aligned with elements of different sub-sentences in the English sentence, this means the sentence pairs are mismatched. If this occurs, we can reject the sentence pair as unsuitable and avoid searching for antecedents for zero pronouns.

Filter	1	2	3	4	Total
Number of times filters were applied	8	214	34	1	256
Number of zero pronouns detected using each filter	18	214	34	2	267

Table 5: Number of sentence pairs processed and zero pronouns detected by each filter

These filters were applied in numerical order from Filter 1 to Filter 4. By applying these filters before applying our rules, we were able to remove some unsuitable sentences successfully, as shown in Table 5. We then reapplied our modified rules, and our new results are shown in Table 6. Due to filtering, the number of rules which were applied during processing was reduced, while also achieving a large increase in the precision of zero pronoun determination (from 67.4% to 92.2%). Recall remains unchanged at 93.2% because the number of correctly determined zero pronouns was unchanged.

	With added rule		With added rule and filters	
Rule	Rules applied correctly/Rules applied (Accuracy)	Zero pronouns determined and linked correctly/Zero pronouns determined (Accuracy)	Rules applied correctly/Rules applied (Accuracy)	Zero pronouns determined and linked correctly/Zero pronouns determined (Accuracy)
1	95/96 (99.0%)	95/96 (99.0%)	95/96 (99.0%)	95/96 (99.0%)
2	10/11 (90.9%)	10/11 (90.9%)	10/11 (90.9%)	10/11 (90.9%)
3	0/0(-)	0/0 (-)	0/0(-)	0/0 (-)
4	471/471 (100%)	471/471 (100%)	471/471 (100%)	471/471 (100%)
5	17/17 (100%)	34/34 (100%)	17/17 (100%)	34/34 (100%)
6	0/1 (0%)	0/1 (0%)	0/1 (0%)	0/1 (0%)
7	35/64 (54.7%)	35/64 (54.7%)	35/59 (59.3%)	35/59 (59.3%)
8	4/5 (80.0%)	4/5 (80.0%)	4/5 (80.0%)	4/5 (80.0%)
9	61/360 (16.9%)	61/371 (16.4%)	61/83 (73.5%)	61/93 (65.6%)
10	0/0 (-)	0/0 (-)	0/0 (-)	0/0 (-)
Totals	693/1025 (67.6%)	710/1053 (67.4%)	693/743 (93.3%)	710/770 (92.2%)

Table 6: Results with added rule vs. results with added rule and filters

In Table 7, we analyze the distribution of different types of zero pronoun antecedents (based on the results of the added rule and filters, and only including correctly identified antecedent results). The majority of zero pronoun sentences are sentences written in a passive voice, which is due to the nature of the corpus. However, we can also see the distribution of explicit antecedents and specific nouns, the latter of which are especially valuable for accurately translating Japanese sentences into English.

Rule	Number of zero pronoun	They/they	It/it	I	We/we	noun	Passive voice
1	95	27	20	0	11	37	0
2	10	0	0	0	0	10	0
4	471	0	0	0	0	0	471
5	34	0	7	0	0	10	17
7	35	1	7	1	4	22	0
8	4	0	1	0	0	3	0
Totals	632	28	35	1	15	82	488

Table 7: Distribution of zero pronoun antecedents determined using selected rules

5.6 Open test

We then tested the proposed method using 200 unknown sentence pairs. We conducted the test twice, using Nakaiwa's original rules and using our new rule with the added filters.

	Original rules		Added rules with filters	
Rule	Rules applied correctly/Rules applied (Accuracy)	Zero pronouns determined and linked correctly/Zero pronouns determined (Accuracy)	Rules applied correctly/Rules applied (Accuracy)	Zero pronouns determined and linked correctly/Zero pronouns determined (Accuracy)
1	15/16 (93.8%)	15/16 (93.8%)	15/15 (100.0%)	15/15 (100.0%)
2	3/3 (100.0%)	3/3 (100.0%)	3/3 (100.0%)	3/3 (100.0%)
3	0/0 (-)	0/0 (-)	0/0 (-)	0/0 (-)
4	53/53 (100.0%)	53/53 (100.0%)	71/71 (100.0%)	71/71 (100.0%)
5	1/12 (8.3%)	23/24 (95.8%)	4/5 (80.0%)	8/10 (80.0%)
6	0/0 (-)	0/0 (-)	0/0 (-)	0/0 (-)
7	3/11 (27.3%)	3/11 (27.3%)	4/10 (40.0%)	4/10 (40.0%)
8	1/3 (33.3%)	1/3 (33.3%)	1/2 (50%)	1/2 (50%)
9	9/105 (8.6%)	9/123 (7.3%)	9/16 (56.2%)	9/19 (47.4%)
10	0/0 (-)	0/0 (-)	0/0 (-)	0/0 (-)
Totals	95/203 (46.8%)	107/233 (45.9%)	107/122 (87.7%)	111/130 (85.4%)

Table 8: Open test results: using original rules and using added rule with filters

The filters removed 67 unsuitable sentences containing 69 zero pronouns. To calculate recall, we counted the number of zero pronouns, as in Table 2. The total number of zero pronouns in each category was 129. Using the original rules, the precision of zero pronoun determination was 45.9% and recall was 107/129 = 82.9%. Using the added rule with filters, the precision of zero pronoun determination was 85.4% and recall was 111/129 = 86.0%. From these results, we can see that the improved method achieved higher precision and recall compared to the original method. Furthermore, our method was shown to be effective for detecting and resolving zero pronouns of Japanese sentences in unknown corpus. By applying this method to many aligned sentence pairs, it will be possible to collect large amounts of data on recognition of antecedents of zero pronouns, and that data can then be used in many natural language processing applications.

6 Conclusion

This paper proposes a method to detect and identify missing pronouns in written Japanese so they can be inserted into English translations. The proposed method uses aligned sentence pairs from a Japanese-English bilingual corpus and open source tools. Experimental results showed that our proposed method achieved a rule application accuracy rate of 93.3% for all sentence pairs and a 92.2% precision rate for zero pronoun

identification in a closed test (Table 6). We also achieved a rule application accuracy rate of 87.7% and a zero pronoun identification rate of 85.4% in an open test (Table 8). The effectiveness of this method will allow us to automatically construct an annotated, Japanese-English corpus which links zero pronouns in Japanese with their antecedents, which will greatly improve the accuracy of machine translations. In the future, we plan to combine our method with a machine learning technique to investigate Japanese anaphora resolution. We also plan to apply our method to Japanese-English bilingual corpora with other genres, allowing us to examine the full range of zero pronoun phenomena, and to Japanese-Chinese bilingual corpora, allowing us to examine the effectiveness of the proposed method with a bilingual corpus in which both languages have many zero pronouns. The effect of using different syntactic and semantic parsers will also be examined.

7 Acknowledgements

This research was carried out while Zhan Dong was an exchange student at Nagoya University, from September 2014 to July 2015. We would like to thank Prof. Kazuya Takeda for his helpful comments regarding this research.

References

[1] Susan P Converse. Pronominal anaphora resolution in Chinese. 2006.

[2] Niyu Ge, John Hale, and Eugene Charniak. A statistical approach to anaphora resolution. In *Proceedings of the sixth workshop on very large corpora*, volume 71, 1998.

[3] Ryu Iida, Kentaro Inui, and Yuji Matsumoto. Exploiting syntactic patterns as clues in zero-anaphora resolution. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 625–632. Association for Computational Linguistics, 2006.

[4] Ryu Iida, Mamoru Komachi, Kentaro Inui, and Yuji Matsumoto. Annotating a Japanese text corpus with predicate-argument and coreference relations. In *Proceedings of the Linguistic Annotation Workshop*, pages 132–139. Association for Computational Linguistics, 2007.

[5] Ryu Iida and Massimo Poesio. A cross-lingual ILP solution to zero anaphora resolution. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 804–813. Association for Computational Linguistics, 2011.

[6] Satoru Ikehara, Masahiro Miyazaki, Akio Yokoo, Satoshi Shirai, Hiromi Nakaiwa, Kentaro Ogura, Yoshifumi Ooyama, and Yoshihiko Hayashi. Nihongo goi taikei–a Japanese lexicon. *Iwanami Shoten*, 5, 1997.

[7] Taku Kudo and Yuji Matsumoto. Fast methods for kernel-based text analysis. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, pages 24–31. Association for Computational Linguistics, 2003.

[8] Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. Applying conditional random fields to Japanese morphological analysis. In *EMNLP*, volume 4, pages 230–237, 2004.

[9] Shalom Lappin and Herbert J Leass. An algorithm for pronominal anaphora resolution. *Computational linguistics*, 20(4):535–561, 1994.

[10] Ruslan Mitkov. *Anaphora resolution*. Routledge, 2014.

[11] Hiromi Nakaiwa. Automatic extraction of rules for anaphora resolution of Japanese zero pronouns in Japanese–English machine translation from aligned sentence pairs. *Machine translation*, 14(3-4):247–279, 1999.

[12] Hiromi Nakaiwa and Satoshi Shirai. Anaphora resolution of Japanese zero pronouns with deictic reference. In *Proceedings of the 16th conference on Computational linguistics-Volume 2*, pages 812–817. Association for Computational Linguistics, 1996.

[13] Franz Josef Och and Hermann Ney. Giza++: Training of statistical translation models, 2000.

[14] Shanheng Zhao and Hwee Tou Ng. Identification and resolution of Chinese zero pronouns: A machine learning approach. In *EMNLP-CoNLL*, volume 2007, pages 541–550, 2007.