Annotating a Japanese Text Corpus with Predicate-Argument and Coreference Relations

Ryu Iida, Mamoru Komachi, Kentaro Inui and Yuji Matsumoto
Graduate School of Information Science,
Nara Institute of Science and Technology
8916-5 Takayama, Ikoma, Nara, 630-0192, Japan
{ryu-i,mamoru-k,inui,matsu}@is.naist.jp

Abstract

In this paper, we discuss how to annotate coreference and predicate-argument relations in Japanese written text. There have been research activities for building Japanese text corpora annotated with coreference and predicate-argument relations as are done in the Kyoto Text Corpus version 4.0 (Kawahara et al., 2002) and the GDA-Tagged Corpus (Hasida, 2005). However, there is still much room for refining their specifications. For this reason, we discuss issues in annotating these two types of relations, and propose a new specification for each. In accordance with the specification, we built a large-scaled annotated corpus, and examined its reliability. As a result of our current work, we have released an annotated corpus named the *NAIST Text Corpus*[1], which is used as the evaluation data set in the coreference and zero-anaphora resolution tasks in Iida et al. (2005) and Iida et al. (2006).

1 Introduction

Coreference resolution and predicate-argument structure analysis has recently been a growing field of research due to the demands from NLP application such as information extraction and machine translation. With the research focus placed on these tasks, the specification of annotating corpora and the data sets used in supervised techniques (Soon et al., 2001; Ng and Cardie, 2002, etc.) have also grown in sophistication.

For English, several annotation schemes have already been proposed for both coreference relation and argument structure, and annotated corpora have been developed accordingly (Hirschman, 1997; Poesio et al., 2004; Doddington et al., 2004). For instance, in the Coreference task on Message Understanding Conference (MUC) and the Entity Detection and Tracking (EDT) task in the Automatic Content Extraction (ACE) program, which is the successor of MUC, the details of specification of annotating coreference relation have been discussed for several years. On the other hand, the specification of predicate-argument structure analysis has mainly been discussed in the context of the CoNLL shared task[2] on the basis of the PropBank (Palmer et al., 2005).

In parallel with these efforts, there have also been research activities for building Japanese text corpora annotated with coreference and predicate-argument relations such as the Kyoto Text Corpus version 4.0 (Kawahara et al., 2002) and the GDA[3]-Tagged Corpus (Hasida, 2005). However, as we discuss in this paper, there is still much room for arguing and refining the specification of such sorts of semantic annotation. In fact, for neither of the above two corpora, the adequacy and reliability of the annotation scheme has been deeply examined.

In this paper, we discuss how to annotate coreference and predicate-argument relations in Japanese

[1] The NAIST Text Corpus is downloadable from http://cl.naist.jp/nldata/corpus/, and it has already been downloaded by 102 unique users.

[2] http://www.lsi.upc.edu/~srlconll/

[3] The Global Document Annotation

Proceedings of the Linguistic Annotation Workshop, pages 132–139,
Prague, June 2007.

text. In Section 2 to Section 4, we examine the annotation issues of coreference, predicate-argument relations, and event-nouns and their argument relations respectively, and define adequate specification of each annotation task. Then, we report the results of actual annotation taking the Kyoto Corpus 3.0 as a starting point. Section 6 discusses the open issues of each annotation task and we conclude in Section 7.

2 Annotating coreference relations

2.1 Approaches to coreference annotation

Coreference annotation in English has been evolving mainly in the context of information extraction. For instance, in the 6th and 7th Message Understanding Conferences (MUC), coreference resolution is treated as a subtask of information extraction[4]. The annotated corpora built in the MUC contain coreference relations between NPs, which are used as a gold standard data set for machine learning-based approaches to coreference resolution by researchers such as Soon et al. (2001) and Ng and Cardie (2002). However, van Deemter and Kibble (1999) claim that the specification of the MUC coreference task guides us to annotate expressions that are not normally considered coreferential, such as appositive relations (e.g. *Julius Caesar*$_i$, *a well-known emperor*$_i$, ...).

In the task of Entity Detection and Tracking (EDT) in the Automatic Content Extraction (ACE) program (Doddington et al., 2004), the successor of MUC, the coreference relations are redefined in terms of two concepts, *mentions* and *entities*, in order to avoid inappropriate co-indexing. In the specification of EDT, mentions are defined as the expressions appearing in the texts, and entities mean the collective set of specific entities referred to by the mentions in the texts. Entities are limited to named entities such as PERSON and ORGANIZATION for adequacy and reliability of annotation. Therefore, the ACE data set has the drawback that not all coreference relations in the text are exhaustively annotated. It is insufficient to resolve only the annotated coreference relations in order to properly analyze a text.

[4] http://www-nlpir.nist.gov/related_projects/muc/proceedings/co_task.html

2.2 Coreference annotated corpora of Japanese

In parallel with these efforts, Japanese corpora have been developed that are annotated with coreference relations, such as the Kyoto Text Corpus version 4.0 (Kawahara et al., 2002) and GDA-Tagged Corpus (Hasida, 2005). Before reviewing these works, we explain the relationship between anaphora and coreference in Japanese, referring to the following examples. In example (1), the pronoun *sore*$_i$ *(it)* points back to *iPod*$_i$, and these two mentions refer to the same entity in the world and thus are considered both anaphoric and coreferential.

(1) *Tom-wa* *iPod*$_i$*-o* *ka-tta* .
Tom-TOP iPod$_i$-ACC buy-PAST PUNC
Tom bought an iPod.
kare-wa *sore*$_i$*-de* *ongaku-o* *ki-ita* .
he-TOP it$_i$-INS music-ACC listen to-PAST PUNC
He listened to music on it.

On the other hand, in example (2), we still see an anaphoric relation between *iPod*$_i$ *(iPod*$_i$*)* and *sore*$_j$ *(it*$_j$*)* and *sore*$_j$ points back to iPod$_i$. However, these two mentions are not coreferential since they refer to different entities in the world.

(2) *Tom-wa* *iPod*$_i$*-o* *ka-tta* .
Tom-TOP iPod$_i$-ACC buy-PAST PUNC
Tom bought an iPod.
Mary-mo *sore*$_j$*-o* *ka-tta* .
Mary-TOP one$_j$-ACC buy-PAST PUNC
Mary also bought one.

As in the above examples, an anaphoric relation can be either coreferential or not. The former case is called an *identity-of-reference anaphora* (*IRA*) and the latter an *identity-of-sense anaphora* (*ISA*) (see Mitkov (2002)). In English the difference between IRA and ISA is clearly expressed by the anaphoric relations formed with 'it' and 'one' respectively. This makes it possible to treat these classes separately. However, in Japanese, no such clear lexical distinction can be drawn. In both the Kyoto Corpus and GDA-Tagged Corpus, there is no discussion in regards to distinction between ISA and IRA, thus it is unclear what types of coreference relations the annotators annotated. To make matters worse, their approaches do not consider whether or not a mention refers to a specific entity like in the EDT task.

2.3 Annotating IRA relations in Japanese

As described in the previous section, conventional specifications in Japanese are not based on a precise definition of coreference relations, resulting in inappropriate annotation. On the other hand, in our specification, we consider two or more mentions as coreferential in case they satisfy the following two conditions:

- The mentions refer to not a generic entity but to a specific entity.
- The relation between the mentions is considered as an IRA relation.

3 Annotating predicate-argument relations

3.1 Labels of predicate-argument relations

One debatable issue in the annotation of predicate-argument relations is what level of abstraction we should label those relations at.

The GDA-Tagged Corpus, for example, adopts a fixed set of somewhat "traditional" semantic roles such as Agent, Theme, and Goal that are defined across verbs. The PropBank (Palmer et al., 2005), on the other hand, defines a set of semantic roles (labeled ARG0, ARG1, and AM-ADV, etc.) for each verb and annotates each sentence in the corpus with those labels as in (3).

(3) [ARGM−TMP *A year earlier*], [ARG0 *the refiner*] [rel *earned*] [ARG1 *$66 million, or $1.19 a share*].

In the FrameNet (Fillmore and Baker, 2000), a specific set of semantic roles is defined for each set of semantically-related verbs called a FrameNet frame. However, there is still only limited consensus on how many kinds of semantic roles should be identified and which linguistic theory we should adopt to define them at least for the Japanese language.

An alternative way of labeling predicate-argument relations is to use syntactic cases as labels. In Japanese, arguments of a verb are marked by a postposition, which functions as a case marker. In sentence (4), for example, the verb *tabe* has two arguments, each of which is marked by a postposition, *ga* or *o*.

(4) *Tom-ga* *ringo-o* *tabe-ru*
Tom-NOM apple-ACC eat-PRES
(Tom eats an apple.)

Labeling predicate-argument relations in terms of syntactic cases has a few more advantages over semantic roles as far as Japanese is concerned:

- Manual annotation of syntactic cases is likely to be more cost-efficient than semantic roles because they are often explicitly marked by case markers. This fact also allows us to avoid the difficulties in defining a label set.
- In Japanese, the mapping from syntactic cases to semantic roles tends to be reasonably straightforward if a semantically rich lexicon of verbs like the VerbNet (Kipper et al., 2000) is available.
- Furthermore, we have not yet found many NLP applications for which the utility of semantic roles is actually demonstrated. One may think of using semantic roles in textual inference as exemplified by, for example, Tatu and Moldovan (2006). However, similar sort of inference may well be realized with syntactic cases as demonstrated in the information extraction and question answering literature.

Taking these respects into account, we choose to label predicate-argument relations in terms of syntactic cases, which follows the annotation scheme adopted in the Kyoto Corpus.

3.2 Syntactic case alternation

Once the level of syntactic cases is chosen for our annotation, another issue immediately arises, alteration of syntactic cases by syntactic transformations such as passivization and causativization. For example, sentence (5) is an example of causativization, where *Mary* causes *Tom*'s eating action.

(5) *Mary-ga* *Tom-ni* *ringo-o* *tabe-saseru*
Mary-NOM Tom-DAT apple-ACC eat-CAUSATIVIZED
(Mary helps Tom eat an apple.)

One way of annotating these arguments is something like (6), where the relations between the causativized predicate *tabe-saseru (to make someone eat)* and its arguments are indicated in terms of surface syntactic cases.

(6) [REL=*tabe-saseru* (eat-CAUSATIVE), GA=*Mary*, NI=*Tom*, O=*ringo* (apple)]

In fact, the Kyoto Corpus adopts this way of labeling.

An alternative way of treating such case alternations is to identify *logical* (or deep) case relations, i.e. the relations between the *base form* of each predicate and its arguments. (7) illustrates how the arguments in sentence (5) are annotated with logical case relations: *Tom* is labeled as the *ga*-case (Nominative) filler of the verb *tabe (to eat)* and *Mary* is labeled as the *Extra-Nominative* (EX-GA) which we newly invent to indicate the Causer of a syntactically causativized clause.

(7) [REL=*tabe-(ru)* (eat), GA=*Tom*, O=*ringo* (apple), EX-GA=*Mary*]

In the NAIST Text Corpus, we choose to this latter way of annotation motivated by such considerations as follows:

- Knowing that, for example, *Tom* is the filler of the *ga*-case (Nominative) of the verb *tabe (to eat)* in (5) is more useful than knowing that *Tom* is the *ni*-case (Dative) of the causativized verb *tabe-saseru (to make someone eat)* for such applications as information extraction.
- The mapping from syntactic cases to semantic roles should be described in terms of logical case relations associated with bare verbs.

3.3 Zero-anaphora

In the PropBank the search space for a given predicate's arguments is limited to the sentence that predicate appears in, because, syntactically, English obligatory arguments are overtly expressed except pro-form (e.g. *John hopes* [PRO *to leave.*]).

In contrast, Japanese is characterized by extensive use of nominal ellipses, called zero-pronouns, which behave like pronouns in English texts. Thus, if an argument is omitted, and an expression corresponding to that argument does not appear in the same sentence, annotators should search for its antecedent outside of the sentence. Furthermore, if an argument is not explicitly mentioned in the text, they need to annotate that relation as "exophoric." In the second sentence of example (8), for instance, the *ga* (Nominative) argument of the predicate *kaeru (go back)* is omitted and refers to *Tom* in the first sentence. The *kara* (Ablative) argument of that predicate is also omitted, however the corresponding argument does not explicitly appear in the text. In such cases, omitted arguments should be considered as "*exophoric*."

(8) *Tom$_i$-wa* *kyo* *gakko-ni* *it-ta* .
Tom$_i$-TOP today school-LOC go-PAST PUNC
Tom went to school today.
(ϕ_i-ga) *($\phi_{exophoric}$-kara)* *kae-tte* *suguni*
ϕ_i-NOM $\phi_{exophoric}$-ABL go back immediately
(ϕ_i-ga) *kouen-ni* *dekake-ta* .
ϕ_i-NOM park-LOC go out-PAST PUNC
He went to the park as soon as he came back from school.

Table 1: Comparison of annotating predicate-argument relations

corpus	label	search space
PropBank	semantic role	intra
GDA Corpus	semantic role	inter, exo
Kyoto Corpus	surface case (voice alternation involved)	intra, inter, exo
NAIST Corpus (our corpus)	logical (deep) case (relation with bare verb)	intra, inter, exo

intra: intra-sentential relations, inter: inter-sentential relations, exo: exophoric relations

To the best of our knowledge, the GDA-Tagged Corpus does not contain intra-sentential zero-anaphoric relations as predicate-argument relations, so it has a serious drawback when used as training data in machine learning approaches.

Unlike coreference between two explicit nouns where only an IRA is possible, the relation between a zero-pronoun and its antecedent can be either IRA or ISA. For example, in example (8), ϕ_i is annotated as having an IRA relation with its antecedent *Tom$_i$*. In contrast, example (9) exhibits an ISA relation between *iPod$_i$* and ϕ_i.

(9) *Tom-wa* *iPod$_i$-o* *ka$_a$-tta* .
Tom-TOP iPod$_i$-ACC buy$_a$-PAST PUNC
Tom bought an iPod.
Mary-mo *(ϕ_i-o)* *ka$_b$-tta* .
Mary-TOP ϕ_i-ACC buy$_b$-PAST PUNC
Mary also bought one.
[REL=*ka-(u)* (buy), GA=*Mary*, O=*iPod$_i$*]

The above examples indicate that predicate-argument annotation in Japanese can potentially be annotated as either an IRA or ISA relation. Note that in Japanese these two relations cannot be explicitly separated by syntactic clues. Thus, in our corpus we annotate them without explicit distinction. It is arguable that separate treatment of IRA and ISA in predicate-argument annotation could be preferable. We consider this issue as a task of future work.

A comparison of the specification is summarized in Table 1.

4 Annotating event-noun-argument relations

Meyers et al. (2004) propose to annotate semantic relations between nouns referring to an event in the context, which we call event-nouns in this paper. They release the NomBank corpus, in which PropBank-style semantic relations are annotated for event-nouns. In (10), for example, the noun "*growth*" refers to an event and "*dividends*" and "*next year*" are annotated as ARG1 (roughly corresponding to the theme role) and ARGM-TMP (temporal adjunct).

(10) *12% growth in dividends next year* [REL=*growth*, ARG1=*in dividends*, ARGM-TMP=*next year*]

Following the PropBank-style annotation, the NomBank also restricts the search space for the arguments of a given event-noun to the sentence in which the event-noun appears. In Japanese, on the other hand, since predicate-argument relations are often zero-anaphoric, this restriction should be relaxed.

4.1 Labels of event-noun-relations

Regarding the choice between semantic roles and syntactic cases, we take the same approach as that for predicate-argument relations, which is also adopted in the Kyoto Corpus. For example, in (11), *akaji*$_i$ *(deficit)* is identified as the *ga* argument of the event-noun *eikyo (influence)*.

(11) *kono boueki akaji$_i$-wa waga kuni-no*
this trade deficit-TOP our country-OF
kyosoryoku$_j$-ni eikyo-o oyobosu
competitiveness-DAT influence-ACC affect
[REL=*eikyo* (influence), GA=*akaji$_i$* (deficit), O=*kyosoryoku$_j$ (competitiveness)*]
The trade deficit affects our competitiveness.

Note that unlike verbal predicates, event-nouns can never be a subject of voice alternation. An event-noun-argument relation is, therefore, necessarily annotated in terms of the relation between the bare verb corresponding to the event-noun and its argument. This is another reason why we consider it reasonable to annotate the logical case relations between bare verbs and their arguments for predicate-argument relations.

4.2 Event-hood

Another issue to be addressed is on the determination of the "event-hood" of noun phrases, i.e. the task of determining whether a given noun refers to an event or not. In Japanese, since neither singular-plural nor definite-indefinite distinction is explicitly marked, event-hood determination tends to be highly context-dependent. In sentence (12), for example, the first occurrence of *denwa (phone-call)*, subscripted with *i*, should be interpreted as *Tom*'s calling event, whereas the second occurrence of the same noun *denwa* should be interpreted as a physical telephone (cellphone).

(12) *kare$_a$-karano denwa$_i$-niyoruto watashi$_b$-wa*
he$_a$-ABL phone-call$_i$ according to I$_b$-NOM
kare-no ie-ni denwa$_j$-o wasure-tarasii
his-OF home-LOC phone$_j$-ACC leave-PAST
According to his phone call, I might have left my cell phone at his home.

To control the quality of event-hood determination, we constrain the range of potential event-nouns from two different points of view, neither of which is explicitly discussed in designing the specifications of the Kyoto Corpus.

First, we impose a POS-based constraint. In our corpus annotation, we consider only verbal nouns (*sahen*-verbs; e.g. *denwa (phone)*) and deverbal nouns (the nominalized forms of verbs; e.g. *furumai (behavior)*) as potential event-nouns. This means that event-nouns that are not associated with a verb, such as *jiko (accident)*, are out of scope of our annotation.

Second, the determination of the event-hood of a noun tends to be obscure when the noun constitutes a compound. In (13), for example, the verbal noun *kensetsu (construction)* constituting a compound *douro-kensetsu (road construction)* can be interpreted as a constructing event. We annotate it as an event and *douro (road)* as the *o* argument.

(13) *(ϕ-ga) douro-kensetsu-o tsuzukeru*
ϕ-NOM road construction-ACC continue
Someone continues road construction.

In (14), on the other hand, since the compound *furansu kakumei (French Revolution)* is a named-entity and is not semantically decomposable, it is not reasonable to consider any sort of predicate-argument-like relations between its constituents *furansu (France)* and *kakumei (revolution)*.

(14) *furansu-kakumei-ga okoru*
French Revolution-NOM take place
The French Revolution took place.

We therefore do not consider constituents of such semantically non-decomposable compounds as a target of annotation.

5 Statistics of the new corpus

Two annotators annotated predicate-argument and coreference relations according to the specifications, using all the documents in Kyoto Text Corpus version 3.0 (containing 38,384 sentences in 2,929 texts) as a target corpus. We have so far annotated predicate-argument relations with only three major cases: *ga* (Nominative), *o* (Accusative) and *ni* (Dative). We decided not to annotate other case relations like *kara*-case (Ablative) because the annotation of those cases was considered even further unreliable at the point where we did not have enough experiences in this annotation task. Annotating other cases is one of our future directions.

The numbers of the annotated predicate-argument relations are shown in Table 2. These relations are categorized into five cases: (a) a predicate and its argument appear in the same phrase, (b) the argument syntactically depends on its predicate or vice versa, (c) the predicate and its argument have an intra-sentential zero-anaphora relation, (d) the predicate and its argument have an inter-sentential zero-anaphora relation and (e) the argument does not explicitly appear in the text (i.e. exophoric). Table 2 shows that in annotation for predicates over 80% of both *o*- and *ni*-arguments were found in dependency relations, while around 60% of *ga*-arguments were in zero-anaphoric relations. In comparison, in the case of event-nouns, *o*- and *ni*-arguments are likely to appear in the same phrase of given event-nouns, and about 80% of *ga*-arguments have zero-anaphoric relations with event-nouns. With respect to the corpus size, we created a large-scaled annotated corpus with predicate-argument and coreference relations. The data size of our corpus along with other corpora is shown in Table 3.

Next, to evaluate the agreement between the two human annotators, 287 randomly selected articles were annotated by both of them. The results are evaluated by calculating recall and precision in which one annotation result is regarded as correct and the other's as the output of system. Note that only the predicates annotated by both annotators are used in calculating recall and precision. For evaluation of coreference relations, we calculated recall and precision based on the MUC score (Vilain et al., 1995). The results are shown in Table 4, where we can see that most annotating work was done with high quality except for the *ni*-argument of event-nouns. The most common source of error was caused by verb alternation, and we will discuss this issue in detail in Section 6. Such investigation of the reliability of annotation has not been reported for either the Kyoto Corpus or the GDA-Tagged Corpus. However, our results also show that each annotating task still leaves room for improvement. We summarize open issues and discuss the future directions in the next section.

Table 3: Data size of each corpus

corpus	size
PropBank I	7,891 sentences
NomBank 0.8	24,311 sentences
ACE (2005 English)	269 articles
GDA Corpus	2,177 articles
Kyoto Corpus	555 articles (5,127 sentences)
NAIST Corpus (ours)	2,929 articles (38,384 sentences)

Table 4: Agreement of annotating each relation

	recall	precision
predicate	0.947 (6512/6880)	0.941 (6512/6920)
ga (NOM)	0.861 (5638/6549)	0.856 (5638/6567)
o (ACC)	0.943 (2447/2595)	0.919 (2447/2664)
ni (DAT)	0.892 (1060/1189)	0.817 (1060/1298)
event-noun	0.905 (1281/1415)	0.810 (1281/1582)
ga (NOM)	0.798 (1038/1300)	0.804 (1038/1291)
o (ACC)	0.893 (469/525)	0.765 (469/613)
ni (DAT)	0.717 (66/92)	0.606 (66/109)
coreference	0.893 (1802/2019)	0.831 (1802/2168)

6 Discussion

6.1 Identification of predicates and event-nouns

Identification of predicates is sometimes unreliable due to the ambiguity between a literal usage and a compound functional usage. For instance, the expression "*to-shi-te*", which includes the verb *shi* (to do), is ambiguous: either the verb *shi* functions as a content word, i.e. an event-denoting word, or it constitutes a multi-word expression together with *to* and *te*. In the latter case, it does not make sense to interpret the verb *shi* to denote an event. However, this judgment is highly context-dependent and we have not been able to devise a reliable criterion for it.

Tsuchiya et al. (2006) have built a functional expression-tagged corpus for automatically classifying these usages. They reported that the agreement ratio of functional expressions is higher than ours. We believe their findings to also become helpful information for annotating predicates in our corpus.

With regards to event-nouns, a similar problem

Table 2: Statistics: annotating predicate-arguments relations

		ga (Nominative)		*o* (Accusative)		*ni* (Dative)	
predicates	(a) in same phrase	177	(0.002)	60	(0.001)	591	(0.027)
106,628	(b) dependency relations	44,402	(0.419)	35,882	(0.835)	18,912	(0.879)
	(c) zero-anaphoric (intra-sentential)	32,270	(0.305)	5,625	(0.131)	1,417	(0.066)
	(d) zero-anaphoric (inter-sentential)	13,181	(0.124)	1,307	(0.030)	542	(0.025)
	(e) exophoric	15,885	(0.150)	96	(0.002)	45	(0.002)
	total	105,915	(1.000)	42,970	(1.000)	21,507	(1.000)
event-nouns	(a) in same phrase	2,195	(0.077)	5,574	(0.506)	846	(0.436)
28,569	(b) dependency relations	4,332	(0.152)	2,890	(0.263)	298	(0.154)
	(c) zero-anaphoric (intra-sentential)	9,222	(0.324)	1,645	(0.149)	586	(0.302)
	(d) zero-anaphoric (inter-sentential)	5,190	(0.183)	854	(0.078)	201	(0.104)
	(e) exophoric	7,525	(0.264)	42	(0.004)	10	(0.005)
	total	28,464	(1.000)	11,005	(1.000)	1,941	(1.000)

also arises. If, for example, a compound noun contains a verbal noun, we have to judge whether the verbal noun can be interpreted as an event-noun or not. Currently, we ask annotators to check if the meaning of a given compound noun can be compositionally decomposed into those of its constituents. However, the judgement of compositionality tends to be highly subjective, causing the degradation of the agreement ratio of event-nouns as shown in Table 4. We are planning to investigate this problem more closely and refine the current compositionality criterion. One option is to build lexical resources of multi-word expressions and compounds.

6.2 Identification of arguments

As we mentioned in 3.1, we use (deep) cases instead of semantic roles as labels of predicate-argument relations. While it has several advantages as discussed in 3.1, this choice has also a drawback that should be removed. The problem arises from lexical verb alternation. It can sometimes be hard for annotators to determine a case frame of a given predicate when verb alternation takes place. For example, sentence (15) can be analyzed simply as in (16a). However, since the verb *shibaru* (bind) has also another alternative case frame as in (16b), the labeling of the case of the argument *kisoku* (rule), i.e. either GA (NOM) or DE (INST) may be undecidable if the argument is omitted.

(15) *kisoku-ga* *hitobito-o* *shibaru*
rule-NOM people-ACC bind
The rule binds people.

(16) a. [REL = *shibaru* (bind), GA = *kisoku* (rule), O = *hitobito* (people)]

b. [REL = *shibaru* (bind), GA = ϕ (exophoric), O = *hitobito* (people), DE (Instrumental) = *kisoku* (rule)]

Similar problems occur for event-nouns as well. For example, the event-noun *hassei (realization)* has both transitive and intransitive readings, which may produce awkward ambiguities.

To avoid this problem, we have two options; one is to predefine the preference in case frames as a convention for annotation and the other is to deal with such alternations based on generic resources of lexical semantics such as Lexical Conceptual Structure (LCS) (Jackendoff, 1990). Creating a Japanese LCS dictionary is another on-going project, so we can collaborate with them in developing the valuable resources.

6.3 Event-hood determination

Event-nouns of some semantic types such as *keiyaku (contract)*, *kisei (regulation)* and *toushi (investment)* are interpreted as either an event or an entity resulting from an event depending on are context. However, it is sometimes difficult to judge whether such an event-noun should be interpreted as an event or a resultant entity even by considering the whole context, which degrades the stability of annotation. This phenomena is also discussed in the NomBank, and we will share their insights and refine our annotation manual in the next step.

6.4 Identification of coreference relation

Even though coreference relation is defined as IRA relations, the lack of agreement on the granularity of noun classes makes the agreement ratio worse. In other words, it is crucial to decide how to annotate abstract nouns in order to improve the annotation.

Annotators judge coreference relations as whether or not abstract nouns refer to the same entity in the world. However, the equivalence of the referents of abstract nouns cannot be reconciled based on real-world existence since by definition abstract nouns have no physical entities in the real world.

As far as predicate-argument relation is concerned, there might be a need for treating generic entities in addition to specific entities as coreferential in some application. For example, one may want to relate *kids* to *children* in sentence (17).

(17) We all want *children* to be fit and healthy. However, the current invasion of fast food is creating overweight and unhealthy *kids*.

The coreference relation between generic nouns are missed in the current specification since we annotate only IRA relations between specific nouns. Even though there are various discussions in the area of semantics, the issue of how to deal with generic nouns as either coreferential or not in real texts is still left open.

7 Conclusion

In this paper, we reported on the current specification of our annotated corpus for coreference resolution and predicate-argument analysis. Taking the previous work of corpus annotation into account, we decided to annotate predicate-argument relations by ISA and IRA relations, and coreference relations according to IRA relations. With the Kyoto Text Corpus version 3.0 as a starting point, we built a large annotated corpus. We also discussed the revelations made from annotating our corpus, and discussed future directions for refining our specifications of the NAIST Text Corpus.

Acknowledgement

This work is partially supported by the Grant-in-Aid for Scientific Research in Priority Areas JAPANESE CORPUS (http://tokuteicorpus.jp).

References

G. Doddington, A. Mitchell, M. Przybocki, L. Ramshaw, S. Strassel, and R. Weischedel. 2004. Automatic content extraction (ace) program - task definitions and performance measures. In *Proceedings of the 4rd International Conference on Language Resources and Evaluation* (*LREC-2004*), pages 837–840.

Charles J. Fillmore and Collin F. Baker. 2000. Framenet: Frame semantics meets the corpus. In *Proceedings of the 74th Annual Meeting of the Linguistic Society of America*.

K. Hasida. 2005. Global document annotation (gda) annotation manual. http://i-content.org/tagman.html.

L. Hirschman. 1997. *MUC-7 coreference task definition*. version 3.0.

R. Iida, K. Inui, and Y. Matsumoto. 2005. Anaphora resolution by antecedent identification followed by anaphoricity determination. *ACM Transactions on Asian Language Information Processing* (*TALIP*), 4:417–434.

R. Iida, K. Inui, and Y. Matsumoto. 2006. Exploiting syntactic patterns as clues in zero-anaphora resolution. In *Proceddings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics* (*COLING-ACL*), pages 625–632.

R. Jackendoff. 1990. *Semantic Structures*. Current Studies in Linguistics 18. The MIT Press.

D. Kawahara, T. Kurohashi, and K. Hasida. 2002. Construction of a japanese relevance-tagged corpus (in japanese). In *Proceedings of the 8th Annual Meeting of the Association for Natural Language Processing*, pages 495–498.

K. Kipper, H. T. Dang, and M. Palmer. 2000. Class-based construction of a verb lexicon. In *Proceedings of the 17th National Conference on Artificial Intelligence and 12th Conference on on Innovative Applications of Artificial Intelligence*, pages 691–696.

A. Meyers, R. Reeves, C. Macleod, R. Szekely, V. Zielinska, B. Young, and R. Grishman. 2004. The nombank project: An interimreport. In *Proceedings of the HLT-NAACL Workshop on Frontiers in Corpus Annotation*.

Ruslan Mitkov. 2002. *Anaphora Resolution*. Studies in Language and Linguistics. Pearson Education.

V. Ng and C. Cardie. 2002. Improving machine learning approaches to coreference resolution. In *Proceedings of the 40th ACL*, pages 104–111.

M. Palmer, D. Gildea, and P. Kingsbury. 2005. The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–106.

M. Poesio, R. Mehta, A. Maroudas, and J. Hitzeman. 2004. Learning to resolve bridging references. In *Proceddings of the 42nd Annual Meeting of the Association for Computational Linguistics* (*ACL*), pages 144–151.

W. M. Soon, H. T. Ng, and D. C. Y. Lim. 2001. A machine learning approach to coreference resolution of noun phrases. *Computational Linguistics*, 27(4):521–544.

M. Tatu and D. Moldovan. 2006. A logic-based semantic approach to recognizing textual entailment. In *Proceddings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics* (*COLING-ACL*), pages 819–826.

M. Tsuchiya, T. Utsuro, S. Matsuyoshi, S. Sato, and S. Nakagawa. 2006. Development and analysis of an example database of japanese compound functional expressions. *IPSJ Journal*, 47(6):1728–1741.

K. van Deemter and R. Kibble. 1999. What is coreference, and what should coreference annotation be? In *Proceedings of the ACL '99 Workshop on Coreference and its applications*, pages 90–96.

M. Vilain, J. Burger, J. Aberdeen, D. Connolly, and L. Hirschman. 1995. A model-theoretic coreference scoring scheme. In *Proceedings of the 6th Message Understanding Conference (MUC-6)*, pages 45–52.