# Annotating Semantic Role Information to Japanese Balanced Corpus

**Koichi Takeuchi Masayuki Ueno**
Okayama University
3-1-1, Tshimanaka
7008530 Okayama
koichi@cl.cs.okayama-u.ac.jp

**Nao Takeuchi**
Freelance
Language Analyst

## Abstract

This paper describes the ongoing work of annotating semantic role information of Japanese verbs to corpus, based on the thesaurus of predicate-argument structure we have been developing. One of the characteristics of the annotated corpus is hierarchical semantic role information that allows us to capture the abstracted level of similar relations of arguments as well as the detailed level of distinguishing arguments from attributes. The annotation is applied on Balanced Corpus of Contemporary Written Japanese (BCCWJ). In this paper we describe the aim of this construction and annotation framework and discuss current quality of annotated semantic role labels.

## 1 Introduction

Describing predicate-argument relations, i.e., semantic role labels and semantic frames of predicates are studied from the view of annotated corpora such as PropBank, FrameNet as well as linguistics (Levin, 1993; Pustejovsky, 1995; Jackendoff, 1990; Baker et al., 1998; Jackendoff, 2003). Annotated corpora are useful for natural language processing to make applications such as information extraction systems and question answering systems (Surdeanu et al., 2010; Banarescu et al., 2013) in NLP.

On the other hand, widely used annotated corpora in Japanese adopted surface case markers as describing predicate-argument relations (Iida et al., 2007; Kawahara et al., 2007; Komachi and Iida, 2011). Since the surface case markers are the same as verb-by-verb annotation, a system of semantic role labels crossing verbs and their lexicon are needed for deeper semantic NLP for Japanese texts.

In this context, we annotate semantic role and case frame information to Balanced Corpus of Contemporary Written Japanese (Maekawa, 2008). The semantic role labels are defined on Japanese verbs and adjectives in a freely accessible lexicon Japanese Predicate Thesaurus that we have developed as a case frame as well as sense repository[1].

In this paper first we describe our predicate thesaurus that defines hierarchical semantic role labels that consists of abstracted 31 types and the detailed 72 types with attributes. Second we describe the annotation framework and discuss the quality of annotated results for each level of semantic roles.

## 2 Predicate Thesaurus

### 2.1 Overview of the predicate-argument thesaurus

The predicate thesaurus (Takeuchi et al., 2010) is developed to construct a structural semantic frame repository that contains an interface to syntax, i.e., case frames of Japanese predicates based on lexical conceptual structure (Jackendoff, 1990; Jackendoff, 2003; Kageyama, 1996).

Since the syntactic clues in Japanese are much less than that of English to designate argument type[2], we defined case frames for the arguments and annotated then to example sentences to collect sentences that correspond with their case frames.

Our lexicon has about 23000 example sentences for about 11900 Japanese predicates (about 9100 verbs, 750 adjectives, 2050 adjectival verbs). All example sentences are annotated with 72 types of semantic role labels and 1084 types of frames based on hierarchically defined on extended LCS

[1] http://pth.cl.cs.okayama-u.ac.jp.

[2] Subjects are often omitted, and case markers in Japanese are highly ambiguous; for example, English prepositions such as *in, at, with* are correspond to one case marker *de* in Japanese.

form.

Since each class is a cluster of predicates, then the lexicon can be used as thesaurus for taking similar predicates according to the granularity. Let's leave the detailed explanations of the lexicon to previous paper (Takeuchi et al., 2010; Takeuchi et al., 2011), we would like to focus on the semantic role labels we defined in the next section.

### 2.2 Semantic role labels

Semantic role labels are defined according to previous work of lexical semantics (e.g, (Kageyama, 1996; Jackendoff, 1990; Jackendoff, 2003; SGWJG, 2009) and so on). Semantic role labels consist of two levels: the first level is a main relation type between a predicate, (e.g., *Agent, Theme, Goal*) and its argument; the second level designates sub-type of relation or selections restriction of its argument.

As the description of the second level, we put the sub-type information in parentheses. Let us introduce an example of the verb *kasu (lend)*.

```
Kare-wa jitensha-wo watashi-ni kashi-ta
he-Nom  bicycle-ACC   me-DAT     lend-PAS
Agent   Theme       Goal(Person)
He lent a bicycle to me.
```

The dative case of *lend* can be a person who gets the thing lent. To express this relation type in a label, we define a semantic role *Goal* that indicates an end point of change-of-state event, and the attribute *Person* that designates what type of argument can take; finally, we describe the semantic role label as *Goal (Person)* using the combination of the role and attribute[3].

The defined sub-types are here: *Person, Location, Time, Body_Part, Emotion, Material, Product, Event, Action, State, Abstraction, Degree, Object_operated_by_person*. These sub-types can be partially combined with the first level role types. Currently we define 31 types of the first level semantic role labels and 72 types of the second level.

## 3 Annotation of Semantic Role Labels to BCCWJ

BCCWJ contains documents in various kinds of genre, i.e., not only newspapers but also white papers, blogs and novels in Japanese, and then several annotation projects such as Japanese FrameNet (Ohara et al., 2011) and dependency parsing by NAIST (Asahara, 2013), case-marker-based predicate-argument and correference annotation (Komachi and Iida, 2011) are currently going on. Since all of the above annotations are done at the core part in BCCWJ, we can compare our tags to the other annotation results on the same documents by annotating our semantic role labels to the core part.

In the following sections we show that the framework of annotation and the results of annotation quality.

### 3.1 Annotation issues

The aim of the annotation to BCCWJ is to expand the example sentences annotated with semantic role labels and frames because Predicate Thesaurus currently has one or two example sentences for semantic frame. Thus we only annotate the predicates registered in the thesaurus. Annotating semantic role labels to sentences requires previously 1) to collect example sentences for predicates in the thesaurus, 2) to identify arguments of target predicates in sentences[4] and 3) to determine a semantic frame for a polysemous predicate[5].

We prepared the example sentences at the step 1) before starting this annotation task, but annotating arguments (i.e., dependency analyses) of each target predicate is to be asked to annotators. The task of argument annotation, i.e., step 2) can be dealt with on not character level but morpheme level because all of the morphemes are manually analyzed in the core part of BCCWJ.

Annotating 72 types of semantic role labels indicates that annotators have to select one correct label from 72 types variety. This task setting seems to be hard but most of the cases are easier task because the annotators can select 4 or 5 ambiguities from reference correctly annotated examples in Predicate Thesaurus, i.e., gold standard corpus. Thus stable annotation can be expected.

The other issue is qualification of annotators. Because of the limitation of the environment, we cannot hire linguistic specialist but can only students, i.e., non-specialists, for the annotation, we have to give enough instruction for the annotators.

[3]This relation is expressed as *Patient* in conventional linguistics study (Levin and Hovav, 2005). In LCS research, Jackendoff (90:22) describes these sub-types as *ontological categories*.

[4]This is because the project of constructing dependency parsed texts in BCCWJ as mentioned above is still undergoing and not published.

[5]Since this paper focuses on the semantic role annotation task, we omit the details of this step.

In the practical work, we took three months for training annotators and then two months for evaluating inter-annotator agreement.

### 3.2 Framework of annotation

The main steps of the annotation task consists of the following three for a predicate: to select f1) arguments, f2) a semantic frame for the target predicate, and f3) a semantic roles.

To help this annotation task we construct a browser-based annotation system (Figure 1) that serves functions of user management, data management of example sentences annotated tags, showing gold standard annotated examples from Predicate Thesaurus.

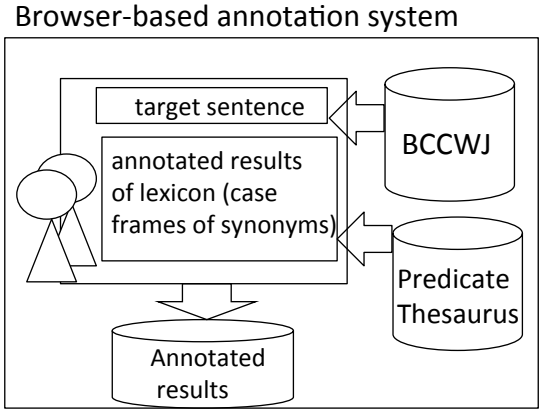


Figure 1: Over view of the annotation system.

Suppose that the semantic frame of the Japanese verb *kau* is decided as *buy*, the annotation system can provide the example sentences with case frames e.g. *kanojo-ha/Agent hon-wo/Theme kau (she buys a book)*. Thus the system can lighten the burden of selection from 72 candidates.

For quick annotation, we hire several annotators. Each annotator takes more than three hours lecture of semantic roles systems. All of the annotators are paired, and thus more than one annotator gives semantic role labels for the same sentence. For the first stage, the target of the predicates are limited to the verbs registered in the thesaurus.

### 3.3 Annotation results and discussions

This is on going research project, however we show the current results of annotation. We hired four annotators who are the students whose domains are in literature, biology, computer science, and law. By the five months work, we finally obtained 7355 instances of verbs (i.e., sentences) on 784 types, and the number of annotated arguments are 12647.

Depending on their different period of employment, we make the pairs of annotators and assign the sentences to the pairs. As described above, the first three months are the training period, then the last two months are used for the following evaluation of the annotators performance. The period of evaluating data is short but the amount of annotated date is about 45% of all annotated sentences.

In the following Table 1 and Table 2 we show that the agreement rates and kappa value of both the detailed (i.e., 72 types) and the normal level (i.e., 31 types) of semantic roles for each working pair.

Table 1: Annotation results for 72 semantic role types.

| pair | #verb | agreement | kappa |
|---|---|---|---|
| A vs. B | 1817 | 0.8190 (2764 / 3375) | 0.7951 |
| A vs. C | 656 | 0.7217 (866 / 1200) | 0.6855 |
| A vs. D | 535 | 0.7914 (831 / 1050) | 0.7668 |
| B vs. C | 497 | 0.7602 (748 / 984) | 0.7318 |
| C vs. D | 146 | 0.75 (228 / 304) | 0.726 |

Table 2: Annotation results of 31 semantic role types.

| pair | agreement | kappa |
|---|---|---|
| A vs. B | 0.8723 (2944 / 3375) | 0.8454 |
| A vs. C | 0.7925 (951 / 1200) | 0.7523 |
| A vs. D | 0.8267 (868 / 1050) | 0.7668 |
| B vs. C | 0.7947 (782 / 984) | 0.7617 |
| C vs. D | 0.7862 (239 / 304) | 0.7604 |

In the both tables, the numerator and the denominator at the agreement column denote the number of semantic labels are agreed for the two annotators, the number of arguments. Since argument positions also decided by the annotators, the number of arguments denotes that the two annotators judged the same argument position. We omitted to show all the detailed statistics of the precisions of recognizing arguments in each sentence, but we show an example of the case at the pair A vs B; the precisions of annotator A and B are 84.5% (3375/3991) and 78.4% (3375/4306), respectively.

As for the kappa value, Table 1 shows high kappa values i.e., more than 0.7 score at most of the pairs. This indicates that the annotation of

72 types categories was done in successfully even though the annotators are not the specialist of linguistics. Comparing Table 1 with Table 2, we find that the normal level of semantic roles are annotated with higher agreement rates and kappa values than the detailed level. The results show the high possibility that the normal level of semantic roles may be easy to be recognized by annotators because of the organized label design.

Currently we are making a gold standard corpus for the annotate semantic roles; using the gold corpus we are planning to evaluate exact performance of the non-specialist annotators work.

## 4 Conclusions

This paper describes an ongoing research work of annotating semantic role labels to a Japanese balanced corpus BCCWJ. The semantic roles are 72 types, but the experimental results of annotation task show that the kappa values are high i.e., from 0.69 to 0.80; this indicates that the detailed semantic role annotation task is promising.

In the future we make a gold standard corpus for the annotate semantic roles, and then we will reveal the possibilities of annotation work of deep semantic tags by non-specialists.

## Acknowledgment

This research received support from JSPS KAKENHI Grant Number 26370485.

## References

M. Asahara. 2013. Memorandum of Dependency and Parallel Phrase Annotation to BCCWJ (Version 0.6). In *Technical Report of Center for Corpus Development, National Institute of Japanese Language and Linguistics*. (in Japanese).

C. F. Baker, C. J. Fillmore, and J. B. Lowe. 1998. The Berkeley FrameNet project. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics*, pages 86–90.

L. Banarescu, C. Bonial, S. Cai, M. Georgescu, K. Griffitt, U. Hermjakob, K. Knight, P. Koehn, M. Palmer, and N. Schneider. 2013. Abstract Meaning Representation for Sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 178–186.

R. Iida, M. Komachi, K. Inui, and Y. Matsumoto. 2007. Annotating a Japanese Text Corpus with a Predicate-Argument and Coreference Relations. In *Proceedings of the 1st Linguistic Annotation Workshop*, pages 132–139.

R. Jackendoff. 1990. *Semantic Structures*. MIT Press.

R. Jackendoff. 2003. *Foundations of Language: Brain, Meaning, Grammar, Evolution*. Oxford University Press.

T. Kageyama. 1996. *Verb Semantics*. Kurosio Publishers. (In Japanese).

D. Kawahara, S. Kurohashi, and K. Hashida. 2007. Construction of a japanese relevance-tagged corpus. In *Proceedings of the 8th Annual Meeting of the Association for Natural Language Processing*, pages 495–498. (In Japanese).

M. Komachi and R. Iida. 2011. Annotating a japanese balanced corpus (bccwj) with a predicate-argument and coreference relations. In *Workshop for Japanese Corpus*, pages 352–330. (In Japanese).

B. Levin and M. R. Hovav. 2005. *Argument Realization*. Cambridge.

B. Levin. 1993. *English Verb Classes and Alternations*. University of Chicago Press.

K. Maekawa. 2008. Balanced corpus of contemporary written japanese. In *Proceedings of the 6th Workshop on Asian Language Resources (ALR)*, pages 101–102.

Study Group of Written Japanese Grammar. 2009. *Contemporary Japanese Grammar 2: Syntax and Voice*. Kurosio Publishers. (in Japanese).

K. Ohara, J. Kato, and H. Saito. 2011. Annotation of Japanese FrameNet to BCCWJ. In *Proceedings of the Workshop of Japanese Corupus in Grant-in-Aid for Scientific Research on Priority Areas*, pages 513–518. (in Japanese).

J. Pustejovsky. 1995. *The Generative Lexicon*. MIT Press.

M. Surdeanu, Massimiliano C., and H. Zaragoza. 2010. Learning to rank answers to non-factoid questions from web collections. *Computational Linguistics*, 37(2):351–383.

K. Takeuchi, K. Inui, N. Takeuchi, and A. Fujita. 2010. A Thesaurus of Predicate-Argument Structure for Japanese Verbs to Deal with Granularity of Verb Meanings. In *The 8th Workshop on Asian Language Resources*, pages 1–8.

K. Takeuchi, S. Tsuchiyama, M. Moriya, Y. Moriyasu, and K. Satoh. 2011. Verb Sense Disambiguation Based on Thesaurus of Predicate-Argument Structure. In *Proc. of the International Conference on Knowledge Engineering and Ontology Development*. 208–213.