Building Japanese Predicate-argument Structure Corpus using Lexical Conceptual Structure

Yuichiroh Matsubayashi, Yusuke Miyao, Akiko Aizawa

National Institute of Informatics
2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo, Japan
{y-matsu, yusuke, aizawa}@nii.ac.jp

Abstract

This paper introduces our study on creating a Japanese corpus that is annotated using semantically-motivated predicate-argument structures. We propose an annotation framework based on Lexical Conceptual Structure (LCS), where semantic roles of arguments are represented through a semantic structure decomposed by several primitive predicates. As a first stage of the project, we extended Jackendoff's LCS theory to increase generality of expression and coverage for verbs frequently appearing in the corpus, and successfully created LCS structures for 60 frequent Japanese predicates in Kyoto university Text Corpus (KTC). In this paper, we report our framework for creating the corpus and the current status of creating an LCS dictionary for Japanese predicates.

Keywords: predicate-argument structure, semantic role labeling, lexical conceptual structure

1. Introduction

This paper introduces an on-going project of creating a Japanese corpus that is annotated using semantically motivated predicate-argument structures.

Due to the success of analyzing syntactic structure of sentences with high accuracy, many researchers pay again much attention to semantic structures inside/outside of a sentence and the semantic relationships between two mentions, such as paraphrasing and textual entailment. One of the key technologies for these problems is predicate-argument structure analysis which detects syntactic/semantic relationships between a predicate and other components in a sentence. In particular, to deal with paraphrasing and textual entailment, relationships between predicates and between their argument structures are required.

For English, FrameNet (Ruppenhofer et al., 2006) and the combination of PropBank (Kingsbury and Palmer, 2002) and VerbNet (Kipper et al., 2000) give such syntactic/semantic relationships between argument structures. However, there are not enough resources in Japanese to obtain such relationships. NAIST text corpus (NTC) (Iida et al., 2007), which is the only corpus including a sufficient amount of analyzed argument structures of Japanese texts, just annotates the tags "ga" (nominative case marker), "wo" (accusative marker) and "ni" (dative marker) each of which corresponds to the *kana* character of the case marker. Thus the annotation is too coarse to capture a semantic function of each argument. Moreover, these three categories are not enough to identify all the core arguments of predicates.

Our purpose is to construct a new resource enriched the information for analyzing predicate-argument structure of Japanese texts. The resource consists of a frame dictionary which gives a complete set of semantically defined arguments for each predicate and texts where predicate-argument structures are annotated based on the dictionary.

[1]Here we omitted the attributes taken by some predicates, in order to simplify the explanation.

[私が/から $_i$]	[それを $_j$]	[彼に $_k$]	伝え ます。
watashi-ga/-kara	sore-wo	kare-ni	tsutaemasu.
I-NOM/-ABL	it-ACC	him-DAT	will tell.

(I will tell it to him.)

伝える (tsutaeru), tell.v

$$\left[\begin{array}{l}\text{cause(affect}(i,j)\text{, go}(j, \left[\begin{array}{l}\text{from(locate(in}(i))) \\ \text{to(locate(at}(k)))\end{array}\right]\text{))} \\ \text{for}\left[\text{cause(affect}(k,j)\text{, go}(j, \left[\text{to(locate(in}(k)))\right]\text{))}\right]\end{array}\right]$$

Figure 1: LCS structure for verb 伝える and its case alternation.[1]

The argument structures in our corpus are based on the theory of Lexical Conceptual Structure (LCS) (Jackendoff, 1990) which represents a semantic structure of a predicate by decomposing its meaning into a combination of several primitive predicates. The motivation in using LCS is to clearly represent a semantic function of each syntactic argument. The primitives in LCS are designed to represent a full or partial action-change-state chain. Each argument slot of the primitive predicates roughly corresponds to a typical thematic role, but highly functionalized and not semantically duplicated (Matsubayashi et al., 2012). In addition, one syntactic argument can be simultaneously filled into different argument slot of primitives in a LCS structure. This gives us a natural understanding for some syntactic alternations in Japanese.[2]

As a first stage of the project, we created a framework for annotating predicate-argument structure based on a LCS theory and developed a LCS dictionary for 60 Japanese verbs by enhancing the theory. We report our framework

[2]For example, in the LCS structure for the verb 伝える in Fig. 1, the argument *i* appears twice: as a first argument of *affect* and as an argument inside of *from*. We found that verbs having this characteristic can alternate a nominative marker *ga* with an ablative marker *kara*.

[私は $_i$]	[彼から $_k$]	[招待を $_j$]	受け た。
watashi-ha	kare-kara	shoutai-wo	uketa.
I-NOM	him-ABL	invitation-ACC	received.

(I received an invitation from him.)

受ける (ukeru), receive.v

$$\left[\begin{array}{l}\text{cause(affect}(i,j)\text{, go}(j, \left[\text{to(locate(in}(i))\right])) \\ \text{comb}\left[\text{cause(affect}(k,j)\text{, go}(j, \left[\begin{array}{l}\text{from(locate(in(k)))} \\ \text{to(locate(at}(i)))\end{array}\right])\right]\end{array}\right]$$

Figure 2: LCS structure for verb 受ける.

for creating the corpus and the current status of creating an LCS dictionary for Japanese predicates.

2. Framework

Similarly to PropBank and FrameNet, we took a framework where a corpus consists of two resources: (1) a frame dictionary which is a set of argument structures for verbs where the arguments have some information of their semantics and (2) annotated texts based on that dictionary. This approach is to give a complete set of syntactic arguments for each predicate and to analyze relationships between syntactic and semantic structures of the arguments. Also, for this reason, our argument-structure annotation using LCS is going to be constructed as another layer on the existing representative corpus (Kyoto university text corpus (Kurohashi and Nagao, 1997); KTC) for Japanese, where syntactic dependency for roughly 40,000 sentences is annotated.

For PropBank and FrameNet, an entry of the dictionary is a sense or concept of a predicate and the entry includes a set of its arguments' labels whose semantics are defined specifically to the sense or concept of that predicate. In our case, an entry of the dictionary is a LCS structure of a predicate and one predicate may have several different LCS structures. We basically followed to Jackendoff's LCS theory, but modified several parts in order to increase the theoretical coverage for various types of predicates appearing in real-world texts. The biggest change is adding a new primitive predicate "*combination*" (*comb* in short) in order to represent multiple sub-events inside of one predicate. This is a simple extension of Jackendoff's predicate *EXCH* to enhance its usage, but essential for creating LCS structures of predicates appearing in actual data. In our development of 60 Japanese predicates (verb and verbal noun) frequently appearing in KTC, 41 out of 109 frames (37.6%) included multiple events. Moreover, some of these frames are difficult to express a correct semantic structure without multiple events. For example, the structure for the verb 受ける in Fig. 2 has three syntactic arguments *i, j* and *k*, but we cannot include *k* in the first formula with a correct interpretation.[3]

In our framework, semantic roles are separated into two large classes. One class contains the roles each of whose semantics is represented through a primitive predicate of LCS. For these roles, annotators assign an argument id (e.g., *i, j, k*) to each argument of a target predicate in texts based on the LCS structures, in stead of assigning a semantic role label. As we mentioned, the interpretation of each argument slot of LCS's primitive predicates roughly corresponds to a typical thematic role. Therefore, similar but more structured semantic role information is assigned to the arguments through an LCS structure.

The other set of roles contains the ones that can generally appear in many verbs and that are not represented by primitive predicates of LCS, such as *time*, *place*, and *manner*. In order to fix this peripheral role inventory, we firstly started with 52 semantic categories for Japanese case markers shown in *Contemporary Japanese Grammar* (group of Japanese descriptive grammar, 2009), deleted roles represented by LCS primitives, combined similar categories into one role and added new roles we observed in KTC during a preliminary analysis, resulting in a total of 24 roles. Table 1 shows a list of our peripheral roles. We designed the peripheral role set as we can annotate the phrases that are semantically related to a target predicate as many as possible.

Fig. 3 illustrates an overview of our framework. During the annotation process, annotators firstly choose a correct LCS structure for a target verb from a dictionary, then assign these two types of roles to the arguments inside a sentence, looking at both the LCS structure and peripheral role inventory. Basically, we focus on syntactically related arguments. Therefore, we currently do not include anaphoric nor coreference relations in our annotation.

3. Text annotation

We are now on the initial stage of the annotation; we created an initial guideline for tagging these roles with 5,000 randomly sampled instances for 50 verbs (100 instances for each verb).[4] Each instance was annotated in parallel by two out of four annotators. We then revised our guideline through a discussion for fixing gold analysis. In order to evaluate a pilot annotation, we performed following steps:

1. We firstly create an annotation guideline observing instances in the corpus.

2. The four annotators trained themselves using 10 verbs. They individually assigning argument structures to 100 instances for each verb. We then discussed together in order to improve the annotation guideline.

3. To each of other 30 verbs, we assigned two out of four annotators and the annotators individually annotated 40 instances for the verb. We then discussed problematic cases to improve the guideline.

4. The assigned two annotators individually annotated extra 60 instances for each of the 30 verbs. We then calculated an inter-annotater agreement using these instances.

[3] If we contain an argument *k* as a *from* argument in the first formula, that means the action of *i* transfers *j* from *k*. However, in reality, *i* just receive *j*, and *k* transfers *j*.

[4] These verbs were selected from the 60 verbs we will mention in Section 4.

Labels	Descriptions	Examples
Time	Time when the event/state occurs	The party will take place [on next Monday].
Time Start	Starting time of the continuous event/state	I read a book [from 10:00 am].
Time End	Ending time of the continuous event/state	I read a book [until 12:00 am].
Duration	Duration time of the continuous event/state	I am keep drawing [during a summer].
Time Limit	Time limit for happening that event/state	I will submit it [by Monday].
Place	Place where the event/state occurs	We met [at a park].
Reference Point	Point at which the event/state occurs (but it is not a time or place)	His history began [from this single picture].
Manner	How the event is performed/progressed	He put it on the table [carefully].
Cause or Reason	Cause or reason which the event occurs with	
Means	Means or instrument that is used for accomplishing the event	I paint a wall [with a brush]. He left office early [because of headache].
Purpose	Purpose for which the event is performed	She took a taxi [to arrive there in time].
Benefactive	Thing which receives benefit of the event	He shouted at a guy angrily [for her].
Counter Value	Counter value for the action	I will make an extra cup of coffee [for $5].
Boundary for Start	Amount that is trigger for starting the event	
Action Direction	Physical direction for patient, which indicates the place where the action occurs	He shot it [from above].
Theme Start	Starting point of theme having certain width	I read a book [from cover] to cover.
Theme End	Ending point of theme having certain width	I read a book from cover [to cover].
Co-actor	Another actor who does the same event	I walked together [with him].
State of Result	Additional adjectival modifier for resulting state of the event	T-shirt is dyed [blue].
Frequency	Frequency of the event	He drinks beer [once in a week].
Iteration Count	Iteration count of the event	He voted [three times].
Amount of Change	Changing amount of theme	It increased [20%].
Assumption	Assumption related to the event/state	I cannot start this [if you don't finish it yet].
Focused Topic	Topic where the event focused	[As for the price], he has a negative opinion to the store.

Table 1: List of 24 peripheral roles

Class	#instances	Strict	Soft
Core	1544	83.48	91.06
Peripheral	480	61.11	64.60

Table 2: Inter-annotator agreement of core and peripheral roles. The number of instances is an average of two annotators. The agreement rate for each role class is a micro average of the rate for each label.

Since our role labels are assigned to phrase-level constituents, we use two types of agreement rate based on strict and soft matching. Each rate is a harmonic mean of:

$$\frac{\text{\#instances to which both two annotators assigned the tag}}{\text{\#instances to which annotator A assigned the tag}},$$

$$\frac{\text{\#instances to which both two annotators assigned the tag}}{\text{\#instances to which annotator B assigned the tag}}.$$

Note that we evaluated this in order to investigate weak points in the current annotation guideline or the label definition, rather than to show the quality of our corpus.
Table 2 shows the result for each role class. The core arguments have a relatively high agreement rate than the peripheral roles. Most of the disagreements for the core argu-

Label	#instances	Strict	Soft
Assumption	18.5	70.27	70.27
Purpose	10.5	38.71	45.16
Cause or Reason	43.5	80.46	82.76
Time	106	68.87	74.05
Boundary for Start	0.5	N/A	N/A
Manner	106	50	56.6
Time Limit	0.5	N/A	N/A
Iteration Count	17.5	80	80
Time End	2	50	50
Place	99.5	71.36	72.36
Means	34.5	37.68	43.48
Amount of Change	10.5	66.67	66.67
Duration	10.5	47.62	57.14
Co-actor	5.5	90.91	90.91
Time Start	8.5	35.29	35.29
Benefactive	0.5	N/A	N/A
Theme Start	0.5	N/A	N/A
Total	480	61.11	64.60

Table 3: Inter-annotator agreement of peripheral roles. The roles which were not assigned in the evaluation data were omitted.

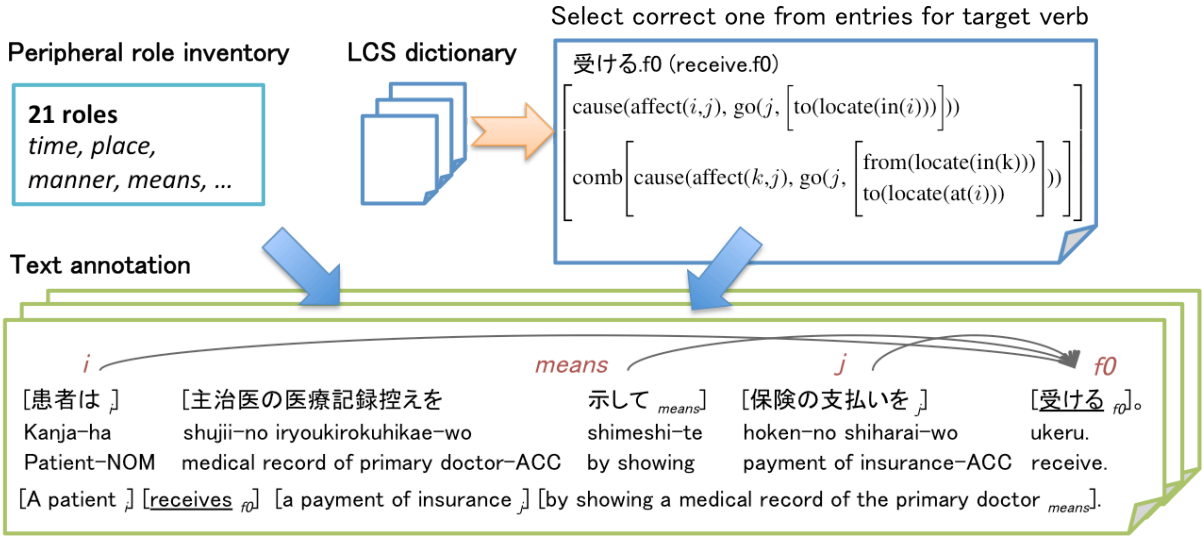

Figure 3: Overview of our framework.

ments are due to careless mistakes of scope, missing case markers, and assignments to coreferential arguments that are not intended to include in our target. The agreement rate for the peripheral roles was lower than 65% even in case of soft matching. Table 3 shows a detailed result for each peripheral role. Some roles including *Cause or Reason*, *Time*, *Place*, *Iteration Count* and *Co-actor* have relatively higher agreements. However, annotators often disagreed on roles such as *Purpose*, *Manner* and *Means*. Most of the disagreements were the case that the other annotator did not assign any labels to the instance. However, *Manner* was often confused with other labels. This is probably because we have not be able to define or explain an exact meaning of *Manner* label.

Finally, we correct the disagreements of these instances and annotated 1,000 instances for other 10 verbs. The goal of the project is to annotate all the predicate-argument relations inside a sentence in KTC and to release it.

4. Building LCS dictionary

Before starting the preliminary annotation, we created LCS structures for 60 verbs (including verbal nouns) frequently appearing in KTC. The purpose here is enhancing LCS theory to cover various types of verbs in real-world data and establishing a consistent way to create a LCS structure for each predicate.

In order to maintain a consistency of the LCS structures among different predicates, we took the following three strategies. First, one of the authors created all the LCS structures alone, looking at the instances of the target verbs in KTC. To increase the coverage of senses and case frames, we also consulted the online Japanese dictionary *Digital Daijisen*[5] and Kyoto university case frames (Kawahara and Kurohashi, 2006) which is a compilation of case frames automatically acquired from a huge web corpus. Second, while doing this step, we developed a decision tree for the first formula in an LCS structure which represents a main event or state focused most in that predicate. This means that we created a finite number of skeleton formulae for a first formula and select one from them using the decision tree. Third, we manually checked a lexical entailment relation between two predicates and tried to construct structures similar to each other. More specifically, we defined several rewriting rules on the LCS structures in order to judge if an LCS structure entails another LCS structure. Intuitively, these rules partially construct a hierarchical graph structure among LCS structures of predicates as shown in Fig. 4, which is similar to the graphical relation of the frames in FrameNet, and thus are useful to check a semantic consistency of LCS structures.

After modifying several parts of LCS theory, we successfully created 109 frames for 60 predicates without any extra modification. The modification was performed to increase the generality of expression by eliminating some primitives that can be expressed by a metaphor for spatial movement. From the result of creating LCS structures, we believe that the resulting theory is stable to some extent. On the other hand, we found that an extra extension of the LCS theory is needed for some verbs. For example, a difference between the case frames for a verb related to reciprocal alteration (see class 2.5 of Levin (Levin, 1993)) such as つながる (connect) and 統一 (integrate) cannot be explained without considering the number of entities in a certain argument. We will continue to create the LCS dictionary by expanding the theory to cover more types of predicates.

5. Conclusion

This paper introduced our study on creating a Japanese corpus that was annotated using semantically-motivated predicate-argument structures. We proposed an LCS-based annotation framework where semantic roles of arguments were represented through a semantic structure decomposed by several primitive predicates. We extended Jackendoff's LCS theory to increase generality of expression and coverage for verbs frequently appearing in the corpus, and suc-

[5] Available at http://dictionary.goo.ne.jp/jn/.

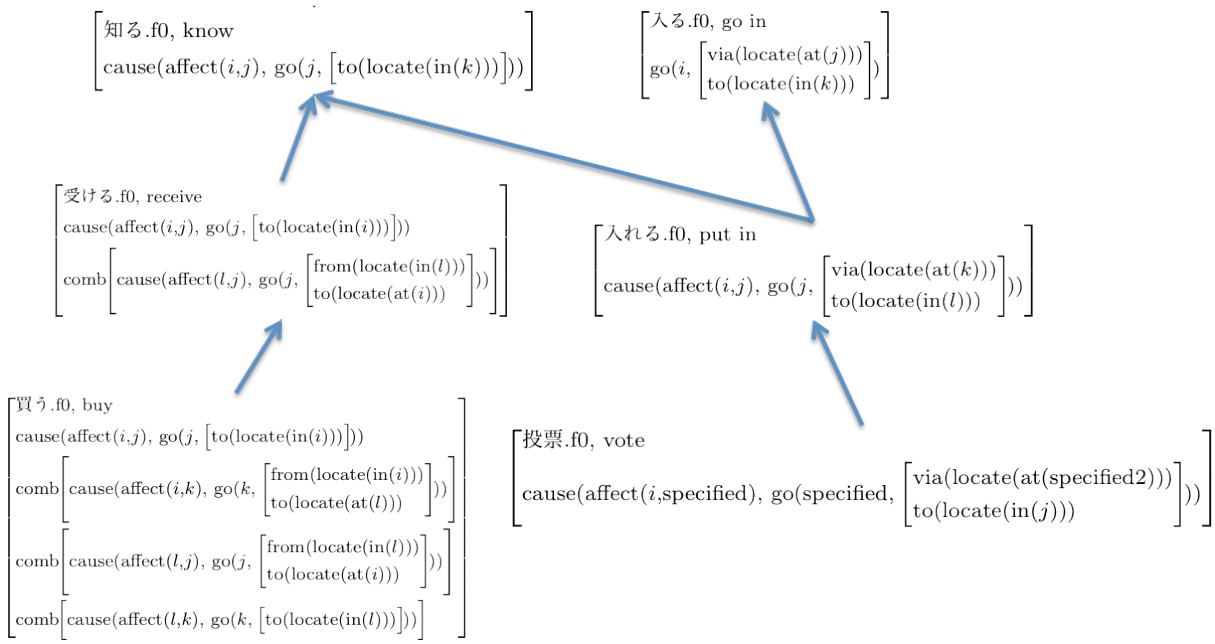

Figure 4: Hierarchical relation of LCS structures created by rewriting rules that we defined on LCS structures. Note that the verb 知る in Japanese is not a stative verb.

cessfully created LCS structures for 60 frequent Japanese predicates in KTC. Our past work mainly focused on creating a theoretical framework for this corpus and now the work is shifting to an actual annotation process. We currently finished annotating 5,000 instances of predicate-argument structures for 50 verbs.

We tried to maintain consistency among LCS structures for semantically related predicates. However, qualitative and quantitative evaluations for our dictionary are necessary future work for this project. To realize this, we plan to formalize a calculation of entailment relations on our LCS structures and compare an automatically generated relation graph with the frame relation in FrameNet.

6. Acknowledgements

This work was partially supported by JSPS Grant-in-Aid for Scientific Research #22800078.

7. References

Study group of Japanese descriptive grammar, editor. 2009. *Contemporary Japanese Grammar*, volume 2. Kuroshio Press.

Ryu Iida, Mamoru Komachi, Kentaro Inui, and Yuji Matsumoto. 2007. 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.

Ray Jackendoff. 1990. *Semantic Structures*. The MIT Press.

D. Kawahara and S. Kurohashi. 2006. Case frame compilation from the web using high-performance computing. In *Proceedings of LREC-2006*, pages 1344–1347.

Paul Kingsbury and Martha Palmer. 2002. From Treebank to PropBank. In *Proceedings of LREC-2002*, pages 1989–1993.

Karin Kipper, Hoa Trang Dang, and Martha Palmer. 2000. Class-based construction of a verb lexicon. In *Proceedings of the National Conference on Artificial Intelligence*, pages 691–696. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999.

Sadao Kurohashi and Makoto Nagao. 1997. Kyoto university text corpus project. *Proceedings of the Annual Conference of JSAI*, 11:58–61.

Beth Levin. 1993. *English verb classes and alternations: A preliminary investigation*. University of Chicago Press.

Yuichiroh Matsubayashi, Yusuke Miyaoy, and Akiko Aizawa. 2012. Framework of semantic role assignment based on extended lexical conceptual structure: Comparison with verbnet and framenet. In *Proceedings of EACL-2012*.

J. Ruppenhofer, M. Ellsworth, M.R.L. Petruck, C.R. Johnson, and J. Scheffczyk. 2006. FrameNet II: Extended Theory and Practice. *Berkeley FrameNet Release*, 1.