Annotation and Analysis of Discourse Relations, Temporal Relations and Multi-layered Situational Relations in Japanese Texts

Kimi Kaneko
Ochanomizu University, Tokyo, Japan
kaneko.kimi@is.ocha.ac.jp

Saku Sugawara
University of Tokyo, Tokyo, Japan
sakus@is.s.u-tokyo.ac.jp

Koji Mineshima
Ochanomizu University, Tokyo, Japan
mineshima.koji@ocha.ac.jp

Daisuke Bekki
Ochanomizu University, Tokyo, Japan
bekki@is.ocha.ac.jp

Abstract

This paper proposes a methodology for building a specialized Japanese data set for recognizing temporal relations and discourse relations. In addition to temporal and discourse relations, multi-layered situational relations that distinguish generic and specific states belonging to different layers in a discourse are annotated. Our methodology has been applied to 170 text fragments taken from Wikinews articles in Japanese. The validity of our methodology is evaluated and analyzed in terms of degree of annotator agreement and frequency of errors.

1 Introduction

Understanding a structured text, such as a newspaper or a narrative, substantially involves the tasks of identifying the events described and locating them in time. Such tasks are crucial for a wide range of NLP applications, including textual entailment recognition, text summarization, and question answering. Accordingly, the task of specifying temporal information in a single text or multiple texts (cross-document event ordering) has been widely used and developed as a temporal evaluation task (Pustejovsky et al., 2009; UzZaman et al., 2012; Minard et al., 2015).

Currently, most work on temporal information processing focuses on relatively simple temporal structures, such as linear timelines. However, understanding the rich temporal content of newspapers and other similar texts often requires accounting for more complex, multi-dimensional information, including not only temporal and causal relations, but also intentional discourse relations (Asher and Lascaridas, 2003).

As an illustration, consider the mini-discourse of Figure 1:

(A) The Independence Day in the United States is annually celebrated on July 4th,
(B) and fireworks shows are held in various parts of the United States at night on that day.
(C) Because my friend invited me to the fireworks show in New York City,
(D) I saw fireworks in Brooklyn Bridge Park on the night of July 4th this year.

Figure 1: Example of discourse units A-B-C-D involving multi-dimensional temporal relations.

In this example, the temporal relation between units (A) and (B), that is, the relation of A-temporally-subsuming-B, can be specified using the temporal expressions *on July 4th* and *at night on that day*; similarly, the relations between (C) and (D), that is, C-temporally-preceding-D, and C-causally-explaining-D, can be specified by the presence of the discourse connective *because*, which explicitly indicates the causal relations.

Beyond these temporal and causal relations, however, a certain kind of temporal relation, as illustrated in the light gray and dark gray squares of Figure 2, occurs between the eventualities (i.e., events or states) described in (A)-(B), on the one hand, and those described in (C)-(D), on the other. A crucial observation is the following: Units (A) and (B) do not describe a specific eventuality (event or state) in a particular past, present or future time, but, instead, describe *general* facts of the entities mentioned (*Independence Day*, etc.); however, units (C) and (D) describe specific events occurring in a particular past time; in particular, (D) introduces an event temporally *subsumed* under the interval described in (B). We say that

Proceedings of the 12th Workshop on Asian Language Resources,
pages 10–19, Osaka, Japan, December 12 2016.

the (A)-(B) sequence describes a situation in the United States at the same general level, whereas the (C)-(D) sequence describes a situation at a specific level; however, (B)-(C) and (B)-(D) shift the layer of the situation from a general to a specific one. Thus, even in a single text, it is crucial to identify multiple levels of a situation described (at a general or a specific level) for a proper understanding of temporal information. We call such a (dis)continuity of a situation or a scene consisting of multiple eventualities (events or states) *a multi-layered situational relation*.

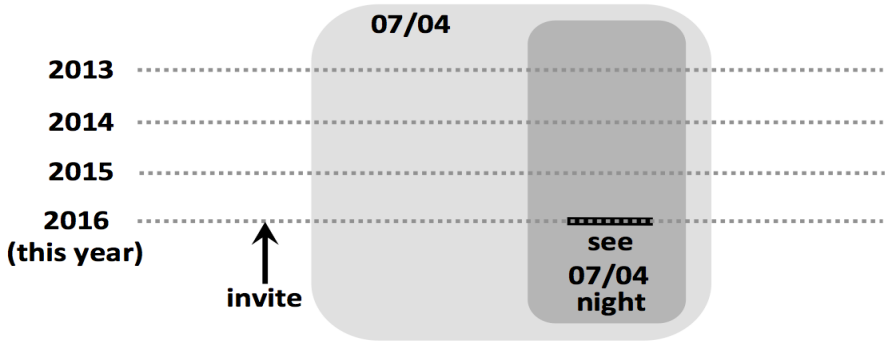

Figure 2: Multi-dimensional temporal information extracted from text in Figure 1.

The primary contribution of this paper is to introduce a new annotation schema refining and enriching previous work on temporal and discourse relation annotation schemata (Asher and Lascaridas, 2003; Kaneko and Bekki, 2014) using multi-dimensional situational relations. On the basis of the proposed method, we report a pilot annotation study of temporal and discourse relations for Japanese news texts, show an evaluation based on degree of inter-annotator agreement, and discuss the results of the annotation experiments and future work.

2 Background and Related Work

In this section, we introduce some existing studies on the annotation of temporal and discourse relations. We hypothesize that some of the difficulties in annotating temporal relations in texts stem from a failure to distinguish between two types of verbal/adjectival expressions in natural language, namely, *individual-level* predicates and *stage-level* predicates, a distinction that has been well-studied in the literature on formal semantics (Krifka et al., 1995). This distinction plays a key role in distinguishing between specific and general levels of situations described in a text. We give an overview of this distinction, which serves as necessary background for the methodology proposed in this paper.

Several specification languages for event and temporal expressions in natural language texts have been proposed, including the annotation specification language TimeML (Pustejovsky et al., 2003a); in addition, annotated corpora, such as TimeBank (Pustejovsky et al., 2003b) and the AQUAINT TimeML Corpus, have been developed. Using TimeML as a base, Asahara et al. (2013) proposed a temporal relation annotation scheme for Japanese and used it to annotate event and temporal expressions in the Balanced Corpus of Contemporary Written Japanese (BCCWJ) (Maekawa et al., 2014). More specifically, based on the framework of TempEval (Pustejovsky et al., 2009; UzZaman et al., 2012; Minard et al., 2015), Asahara et al. (2013) limited target pairs that were annotated temporal relations to the following four types of relations: (i) DCT: relations between a document creation time and an event instance, (ii) T2E: relations between temporal expressions and an event instance within one sentence, (iii) E2E: relations between two consecutive event instances, and (iv) MAT: relations between two consecutive matrix verbs of event instances. They classified event expressions into seven types, including OCCURRENCE and STATE, with respect to which the annotation agreement rates were calculated. They reported that among the seven types of event instances, those pairs containing an expression classified as STATE showed much lower degrees of inter-annotator agreement (0.424) than relations between other event instances. They argued that this difficulty was because recognition of the time interval boundaries for state expressions

was relatively difficult for annotators.

We hypothesize that the difficulty in recognizing time interval boundaries of states (start and end points of states) stems from the fact that the term "state" has the following two senses: (i) permanent/stable properties of individuals and (ii) transient/episodic states applying to a particular stage of an individual. The distinction between (i) and (ii) has long been noticed in the linguistics literature; a predicate expressing a permanent/stable property of an individual is called an *individual-level predicate*, while that expressing a transient/episodic state applying to a particular stage of an individual is called a *stage-level predicate* (Carlson, 1977; Milsark, 1979; Krifka et al., 1995; Kratzer, 1995; Fernald, 2000; Ogawa, 2001). Note here that a predicate expressing a temporal and episodic event is also classified as a stage-level predicate.

For example, (1a), (1b), and (1c) are sentences containing an individual-level predicate (*being a professor of mathematics*), a stage-level predicate for an event (*gave a lecture*), and a stage-level predicate for a state (*was standing during the lecture*), respectively.

(1) a. Susan is a professor of mathematics. INDIVIDUAL-LEVEL/STABLE STATE
b. Today she gave a lecture to her students on geometry. STAGE-LEVEL/EPISODIC EVENT
c. She was standing during the lecture. STAGE-LEVEL/EPISODIC STATE

It seems that those examples containing an individual-level predicate cause the most difficulty in time-interval boundary recognition. For instance, it would be difficult to determine the start and end points for *being a professor of mathematics* in (1a) on the basis of the text; although it is meaningful to ask when Susan became a professor of mathematics, the information about such a temporal boundary is not the main point of statement (1a). Using the terminology introduced in Section 1, (1a) does not describe a specific eventuality (event or state), but states a general state (property) of Susan. In contrast, (1b) and (1c) introduce a temporal event or state with specific temporal boundaries. Thus, (1b) and (1c) report a continuous situation consisting of temporal events and states, while (1a) is a comment, on the individual appearing in that situation, from a different level; that is, a level that is distinguished from the level of the situation described.

It has been noticed in the literature that the distinction between individual-level predicates and stage-level predicates depends on the context of use (McNally, 1998; Jäger, 2001). In the following examples, the predicate *is an olympic swimmer* is used to deliver a temporal and transient state in (2a) extending to (2b), whereas in (3a) it expresses a stable property of John providing background information for understanding (3b).

(2) a. John is an olympic swimmer.
b. He will retire this spring and take up the post of head coach of the junior team.

(3) a. John is an olympic swimmer.
b. He participated in this olympics and was awarded a gold medal.

This means that whether a given predicate is interpreted as individual-level or stage-level often cannot be determined without reference to the surrounding context.

This example also suggests that discourse relations (rhetorical relations), such as BACKGROUND and NARRATION, play a crucial role in determining the distinction between individual-level and stage-level interpretations of predicates (that is, the layer of a situation in our terms) and, for that matter, in determining temporal relations between events/states.

With regard to discourse relations, various theories and specification languages have been proposed in the literature, including Rhetorical Structure Theory (RST) (Mann and Thompson, 1987), Segmented Discourse Representation Theory (SDRT) (Asher and Lascaridas, 2003), and many others (Carlson et al., 2001; Polanyi et al., 2004; Baldridge et al., 2007; Kaneko and Bekki, 2014). Also, annotated corpora based on them have been released, including, most notably, the Penn Discourse TreeBank (PDTB) (Prasad et al., 2005). To our knowledge, however, no label set has been proposed so far that makes a connection between discourse relations and individual/stage-level distinctions and thereby takes into account the relationship between temporal relations and discourse relations.

In fact, the difference in discourse interpretation resulting from the use of individual-level and stage-level predicates is not described by these previous theories of discourse relations. For instance, theories such as RST (Mann and Thompson, 1987) and SDRT (Asher and Lascaridas, 2003) use the discourse relation BACKGROUND to describe the relation between an event description and a state description. However, such an account fails to describe the difference exemplified in (2) and (3) because, in both cases, the first sentence describes a state in the standard sense, whereas the second sentence introduces a set of events.

PDTB (Prasad et al., 2005; Prasad et al., 2014) adopts a *lexically grounded* annotation method, in which annotators are asked to examine lexical items explicitly signaling discourse relations; when such a lexical item is absent, but a particular discourse relation is inferable for adjacent sentences, annotators are asked to find a lexical item that could serve as an explicit signal for the corresponding discourse relation. A particular label (ENTREL) is annotated when no explicit or implicit lexical item is found for adjacent sentences, but the second sentence serves to provide some further description of an entity mentioned in the first sentence (cf. *entity-based* coherence in Knott et al., 2001). This ENTREL label is the majority class label in PDTB. However, similarly to RST and SDRT, PDTB fails to capture the difference exemplified in (2) and (3), since in both examples, the second sentence provides further information about the entity (*John*) in the first sentence.

The ultimate objective of this work is to combine discourse relations, temporal relations, and multi-layered situations triggered by different types of predicates (stage-level and individual-level) in text, and, thereby, to improve existing annotation schemata for discourse and temporal information. We analyze how these different dimensions interact with one another by conducting annotation experiments.

3 Annotation Schema

We present a methodology for annotating discourse relations, temporal relations, and multi-layered situational relations. We limit target pairs for which discourse relations are annotated to (i) main and subordinate clauses in a single sentence and (ii) two consecutive sentences. For temporal relations and multi-layered situational relations, the pair of propositions in each unit is also annotated. By a proposition, we mean a tensed predicate (e.g., *hold*, *invite*, and *see* in Figure 1) denoting either an event or a (generic or specific) state. In the case of a discourse unit consisting of several propositions, such as a complex sentence, we focus on the proposition in the main clause.

The result of annotating the sample text in Figure 1 is shown below.

A-B : [NARRATION(A, B), SUBSUMPTION(A, B), SAME_SITU(A, B)]
B-C : [BACKGROUND(B, C), PRECEDENCE(C, B), SUBSUMPTION_SITU(B, C)]
B-D : [BACKGROUND(B, D), SUBSUMPTION$(B, D$ SUBSUMPTION_SITU(B, D)]
C-D : [EXPLANATION(C, D), PRECEDENCE(C, D), SAME_SITU(C, D)]

Figure 3: Result of tagging text in Figure 1.

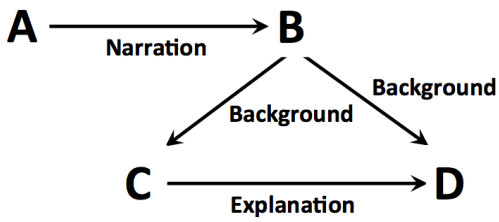

Figure 4: Corresponding discourse graph.

In Figure 3, for each pair (X, Y) of discourse units, we annotate a triple of relations X-Y : $[D, T, S]$, where D is a discourse relation, T is a temporal relation, and S is a multi-layered situational relation between X and Y. These relations are annotated for each pair of discourse units from (A) to (D) in Figure 1. Figure 4 depicts a corresponding discourse graph that indicates the discourse relations and multi-layered situations in Figure 3. Discourse units belonging to the same layer (A-B and C-D) are positioned vertically, whereas those belonging to different layers (B-C and B-D) are positioned horizontally.

The remainder of this section is structured as follows. In Sections 3.1 and 3.2, we deal with temporal relations and multi-layered situational relations, respectively. In Section 3.3, we introduce discourse relations, and describe constraints that these three types of relations impose on one another.

Label		Description	BCCWJ-TimeBank's Labels
TEMP_REL(A, B)	PRECEDENCE(A, B)	end time (A) $<$ start time (B) In other words, eventuality A temporally precedes eventuality B.	before, after, meets, met_by
	OVERLAP(A, B)	start time (A) $<$ start time (B) $<$ end time (A) $<$ end time (B) In other words, eventuality A temporally overlaps with eventuality B.	overlapped_by, overlaps
	SUBSUMPTION(A, B)	start time (A) $<$ start time (B) & end time (B) $<$ end time (A) In other words, eventuality A temporally subsumes eventuality B. Either start times or end times between two eventualities may be simultaneous.	finishes, finished-by, during/is_included, starts, started-by, contains/includes
	SIMULTANEOUS(A, B)	Start time (A) = start time(B) & end time (B) = end time (A) In other words, eventuality A is simultaneous with eventuality B.	equal/identity
NO_TEMP_REL(A, B)		There is no temporal relation between eventuality A and eventuality B.	vague

Table 1: Temporal relations and their correspondence to temporal relations in BCCWJ-TimeBank.

Label	Description	Example
SUBSUMPTION_SITU(A, B)	The layer of situation in which A holds is more general than the one in which B holds.	**A**: The Olympic Games are held every four years. **B**: Tom participated in this Olympic Game.
SAME_SITU(A, B)	A and B hold in the same situational layer. A pair of specific eventualities, or a pair of propositions acting as individual-level predicates.	**A**: I went to the university. **B**: I took a class.

Table 2: Multi-layered situational relations.

3.1 Temporal Relations

On the basis of TimeML (Pustejovsky et al., 2003a) and BCCWJ-TimeBank (Asahara et al., 2013), we use temporal relations: PRECEDENCE, OVERLAP, SUBSUMPTION, and SIMULTANEOUS. When no temporal relations are found, NO_TEMP_REL is annotated. When any of the temporal relations (PRECEDENCE, OVERLAP, SUBSUMPTION, or SIMULTANEOUS) applies, but temporal relations are underspecified, TEMP_REL is annotated. Table 1 summarizes definitions of the temporal relations, and shows their correspondence to BCCWJ-TimeBank temporal relations. Each temporal relation can be defined as a relation between the start time and the end time of two eventualities. We assume that, for all eventualities, the start time of an eventuality e is earlier than its end time.

For each temporal relation, the order of arguments A and B can be reversed; thus, for propositions A and B with which a temporal relation is to be annotated, each temporal relation allows two possibilities; for example, PRECEDENCE(A, B) and PRECEDENCE(B, A). On the basis of these assumptions, the temporal locations of two events described by BCCWJ-TimeBank temporal relations can be reduced to the ones summarized in Table 1.

3.2 Multi-Layered Situational Relations

On the basis of the distinction between individual-level predicates and stage-level predicates as discussed in Section 2, we define multi-layered situational relations as relative differences between layers describing situations. The definition is summarized in Table 2.

For a pair of propositions A and B, multi-layered situational relations are classified into two types. First, SUBSUMPTION_SITU(A, B) indicates that A describes an individual-level, generic situation, whereas B describes a stage-level, more specific situation; hence, they belong to different layers. More specifically, we determine that the relation SUBSUMPTION_SITU(A, B) holds if (i) the main predicate in the proposition A is an individual-level predicate describing a generic state, including stable properties of individuals, and (ii) the main predicate in proposition B is a stage-level predicate describing a more specific situation (event or state). In most cases, the generic state (situation) described in A serves as background knowledge for understanding B. The multi-layered situational relations annotated in Figure 3 contain two instances of this relation, SUBSUMPTION_SITU(B, C) and SUBSUMPTION_SITU(B, D).

Secondly, the relation SAME_SITU(A, B) indicates that eventualities described in A and B belong to the same layer. There are two possibilities: Both A and B describe a stage-level, specific situation, or both A and B describe an individual-level, generic situation.

For tests distinguishing between individual-level and stage-level predicates in a given context, we use

Label	Description	Typical connectives
ALTERNATION(A, B)	"A or B": A and B denote alternative situations.	または（or）
BACKGROUND(A, B)	B describes the background situation of A.	そのとき（then）
CHANGE(A, B)	"A. By the way, B": Relation for switching a topic.	ところで、さて
CONSEQUENCE(A, B)	"If A then B": A is a condition of B.	ならば（if ∼ then ...）
CONTRAST(A, B)	"A but B": B contrasts with A.	しかし（but）
ELABORATION(A, B)	B describes a part of A in detail.	–
EXPLANATION(A, B)	A is a cause, and B is its effect.	ので、から（because）
NARRATION(A, B)	A and B occur (Alternatively, are described) in sequence, and have a common topic. A and B hold in the same situational layer.	そして、それから（and）
INSTANCE(A, B)	"A; for example, B": B describes an instance of A.	例えば（for example）
PARALLEL(A, B)	A and B have similar semantic structures, such as "It is hot in summer. It is cold in winter." Alternatively, an A is simultaneous with B.	同時に（at the same time） かつ（and）
RESTATEMENT(A, B)	B is a paraphrase of A.	つまり（namely）

Table 3: Discourse relations.

two linguistic clues/tests proposed in the literature (Kageyama, 2006). The first clue concerns the type of predicates: The following predicates (typically, appearing in the simple present tense) tend to be interpreted as individual-level predicates (Carlson, 1977).

(4) a. Stative verbs, such as *know*, *love*, *hate*, etc. (cf. *hit*, *run*, etc.)
b. Predicative, post-copular NPs, such as *be a professor* and *be an Olympic athlete*
c. Adjectives, such as *intelligent*, *tall*, *blue*, etc. (cf. *drunk*, *available*, etc.)

Secondly, a stage-level predicate can be modified by an adverbial expression, such as *in a hurry*; a locative modifier, such as *in the car*; or a temporal modifier, such as *just for now* or *today*; whereas an individual predicate cannot (Kratzer, 1995). Thus, the following sentences, understood in a normal context, are anomalous:

(5) a. *Susan is a professor {in a hurry, in the car}.
b. *John knows Latin {in his office, today}.

In addition to the information provided by discourse relations introduced in the next subsection, these linguistic tests and clues are used to distinguish between individual-level (generic/stable) states and stage-level (specific/transient) states.

3.3 Discourse Relations

On the basis of the labels for discourse relations proposed in Kaneko and Bekki (2014), which draw on the classifications in PDTB (Prasad et al., 2005) and SDRT (Asher and Lascaridas, 2003), we use discourse relations, as summarized in Table 3. See Kaneko and Bekki (2014) and Asher and Lascaridas (2003) for more details on the definition of each discourse relation.

As mentioned in Sections 1 and 2, there is a set of discourse relations imposing constraints on temporal relations and multi-layered situational relations. Table 4 shows the manner in which temporal relations, multi-layered situational relations, and discourse relations constrain one another. By annotating discourse relations together with multi-layered situational relations, we can narrow down the range of candidates for temporal relations to be annotated. Correspondences between our labels and those presented in Kaneko and Bekki (2014) and SDRT (Asher and Lascaridas, 2003) are also shown in Table 4.

4 Results and Discussion

We applied our methodology to 90 sentences from Japanese Wikinews articles[2] in June and July 2016. The sentences were decomposed by one annotator, and labels were assigned to the decomposed segments

[2]https://ja.wikinews.org

Our Discourse Relation	Multi-layered Situational Relation Restriction	Temporal Restriction	Discourse Relation in Kaneko and Bekki (2014)	Discourse Relation in SDRT
ALTERNATION(A, B)	–	–	ALTERNATION(A, B)	ALTERNATION(A, B)
BACKGROUND(A, B)	SAME_SITU(A, B)	SUBSUMPTION(A, B)	BACKGROUND(A, B)	BACKGROUND(A, B)
	SUBSUMPTION_SITU(A, B)	–	COMMENTARY(A, B)	COMMENTARY(A, B)
CONSEQUENCE(A, B)	SAME_SITU(A, B)	TEMP_REL(A, B)	CONSEQUENCE(A, B)	CONSEQUENCE(A, B)
CONTRAST(A, B)	–	–	CONTRAST(A, B)	CONTRAST(A, B)
ELABORATION(A, B)	SAME_SITU(A, B)	SUBSUMPTION(A, B)	ELABORATION(A, B)	ELABORATION(A, B)
EXPLANATION(A, B)	SAME_SITU(A, B)	TEMP_REL(A, B)	EXPLANATION(A, B)	EXPLANATION(A, B) RESULT(A, B)
NARRATION(A, B)	SAME_SITU(A, B)	TEMP_REL(A, B)	NARRATION(A, B) ADDITION(A, B)	NARRATION(A, B)
CHANGE(A, B)	–	–	INTRODUCTION(A, B)	NARRATION(A, B)
INSTANCE(A, B)	SUBSUMPTION_SITU(A, B)	–	INSTANCE(A, B)	–
PARALLEL(A, B)	SAME_SITU(A, B)	–	PARALLEL(A, B)	PARALLEL(A, B)
RESTATEMENT(A, B)	SAME_SITU(A, B)	–	COMMENTARY(A, B)	COMMENTARY(A, B)

Table 4: Restrictions that types of relations impose on one another, and correspondences between our methodology, Kaneko and Bekki (2014), and SDRT.

by two annotators. We used the labels presented in Section 3, and assigned "unknown" in cases where pairs could not be labeled. The agreement for 170 pairs generated from 90 pairs and their corresponding Kappa coefficients are presented in Table 5.

Label type	Agreement	Kappa coefficient
Discourse relations	0.69	0.56
Temporal relations	0.74	0.35
Multi-layered situational relations	0.91	0.49
Mean	0.78	0.48
Total	0.89	0.86

Table 5: Agreement and Kappa coefficients in annotations.

The agreement was computed as follows:

$$Agreement = Matching\ labels/Total\ labels$$

Kaneko and Bekki (2014), which used the same set of discourse relations as ours, reported an agreement rate of 0.67 and a Kappa coefficient of 0.57 for discourse relations. Since they computed the agreement by using annotated sentence data, their results are not directly comparable with ours. Nevertheless, the similarity of the values suggests that our method is comparable to that in Kaneko and Bekki (2014) in terms of agreement.

Table 6 shows the distribution of labels for segments in our study, and compares it with that presented in Kaneko and Bekki (2014). We can see from Table 6 that NARRATION was assigned most frequently, both in our study and in Kaneko and Bekki (2014). The number of assignments of SUBSUMPTION_SITU by two annotators showed that they judged that there were some points in texts in which the situation layer had been switched.

The number of pairs for which labels tagged by two annotators were different was 52 for discourse relations, 44 for temporal relations, and 17 for multi-layered situational relations. Table 7 shows the error distribution in this annotation experiment.

Of the 52 pairs for which the two annotators assigned different discourse relations, BACKGROUND and NARRATION were assigned to 14 pairs, NARRATION and PARALLEL to 7 pairs, and NARRATION and EXPLANATION to 7 pairs. One reason that the two annotators assigned different annotations was that we did not impose constraints on BACKGROUND, NARRATION or PARALLEL with respect to assignment of temporal relations and situational relations. These three relations have been known to be difficult

[3]The distribution of labels in Kaneko and Bekki (2014) has been computed on the basis of Table 4.

Label	Segments	
	Kaneko and Bekki (2014)	Ours
ALTERNATION	0	$0 \cap 1 = 0$
BACKGROUND	7	$24 \cap 29 = 19$
CHANGE	8	$7 \cap 1 = 1$
CONSEQUENCE	2	$1 \cap 2 = 1$
CONTRAST	6	$12 \cap 14 = 12$
ELABORATION	23	$8 \cap 5 = 3$
EXPLANATION	10	$20 \cap 15 = 13$
NARRATION	69	$89 \cap 80 = 65$
INSTANCE	6	$0 \cap 0 = 0$
PARALLEL	0	$7 \cap 9 = 4$
RESTATEMENT	–	$0 \cap 0 = 0$
UNKNOWN	–	$0 \cap 1 = 0$
total	**128**	**170**

Annotator 1 $\cap$ Annotator 2 = Match count

Label	Segments
	Ours
TEMP_REL	$14 \cap 57 = 7$
PRECEDENCE	$88 \cap 60 = 45$
OVERLAP	$7 \cap 1 = 0$
SUBSUMPTION	$28 \cap 38 = 16$
SIMULTANEOUS	$30 \cap 3 = 3$
NO_TEMP_REL	$2 \cap 7 = 0$
UNKNOWN	$0 \cap 3 = 0$
total	**170**
SUBSUMPTION_SITU	$18 \cap 17 = 10$
SAME_SITU	$150 \cap 152 = 143$
UNKNOWN	$0 \cap 1 = 0$
total	**170**

Annotator 1 $\cap$ Annotator 2 = Match count

Table 6: Distribution of labels for segments in Kaneko and Bekki (2014) and in our study[3].

Annotator-1's label	Annotator-2's label	Frequency	Annotator-1's label	Annotator-2's label	Frequency
Discourse relation			Others		
BACKGOUND	NARRATION	14	PRECEDENCE	SUBSUMPTION	3
			SIMULTANEOUS	SUBSUMPTION	1
			SUBSUMPTION	NO_TEMP_REL	1
			SUBSUMPTION_SITU	SAME_SITU	4
PARALLEL	NARRATION	7	SIMULTANEOUS	NO_TEMP_REL	1
			PRECEDENCE	SUBSUMPTION	1
EXPLANATION	NARRATION	7	–		
BACKGOUND	ELABORATION	6	SIMULTANEOUS	SUBSUMPTION	2
			SUBSUMPTION_SITU	SAME_SITU	1
CHANGE	NARRATION	4	PRECEDENCE	SUBSUMPTION	1
			SUBSUMPTION_SITU	SAME_SITU	1
temporal relation			Others		
PRECEDENCE	SUBSUMPTION	13	BACKGOUND	NARRATION	3
			SUBSUMPTION_SITU	SAME_SITU	1
SIMULTANEOUS	SUBSUMPTION	7	BACKGOUND	ELABORATION	2
			SUBSUMPTION_SITU	SAME_SITU	2
PRECEDENCE	OVERLAP	6	–		
Multi-layered situational relation			Others		
SUBSUMPTION_SITU	SAME_SITU	15	BACKGOUND	NARRATION	4
			TEMP_REL	NO_TEMP_REL	4

Table 7: Error distribution in annotation exercise (excerpted).

to distinguish by use of a test involving insertion of a lexical item, which was used in the annotation schema of PDTB. Thus, it seems necessary to define temporal and situation constraints more precisely, or to introduce label sets for which any insertion test would be applicable.

Regarding temporal relations for which the two annotators assigned different labels, PRECEDENCE and SUBSUMPTION were assigned to 13 pairs, SIMULTANEOUS and SUBSUMPTION to 7 pairs, and PRECEDENCE and OVERLAP to 6 pairs. There are several possible reasons for these discrepancies. First, these seem to be cases in which we cannot precisely recognize time intervals, such as (B) and (D) in Figure 1; in this case, (B) and (D) only contain temporal information for *on the night of July 4th*, and therefore, SIMULTANEOUS can be assigned to this pair, as well as SUBSUMPTION. In addition, for the 6 pairs that had labeling inconsistencies between PRECEDENCE and OVERLAP, the two annotators labeled the same discourse relations and the same multi-layered situational relations. With these points in mind, our methodology should reflect partial (in)consistencies of decision, such as "we can only determine that the two eventualities temporally overlap, although their start and end point are unknown" or "we can only

determine the order between the starting points of the two eventualities, although the exact time intervals of the two eventualities are ambiguous."

For multi-layered situational relations, 15 pairs were assigned SUBSUMPTION and SAME_SITU. These errors were mainly caused by ambiguity in the examples and lack of constraints imposed on discourse relations and temporal relations, as shown in Table 4. A refinement of constraints is necessary to improve the quality of annotation.

5 Conclusion

This paper proposed a methodology for building a specialized Japanese dataset for recognizing temporal relations and discourse relations. We introduced multi-layered situational relations triggered by distinctions between individual-level and stage-level predicates in text, as well as constraints imposed by each type of relation. We conducted annotation experiments in which we applied our methodology to 170 pairs of text fragments from Japanese Wikinews articles. We compared our method with that of Kaneko and Bekki (2014) in terms of agreement. In future work, we intend to address the issues discussed in Section 4. We also plan to build an inference model suited for the methodology presented in this work.

Acknowledgements

We would like to thank the anonymous reviewers at ALR12 for insightful comments. This work was supported by a Grant-in-Aid for JSPS Research Fellows, Grant Number 15J11737 and the JST CREST program, Establishment of Knowledge-Intensive Structural Natural Language Processing and Construction of Knowledge Infrastructure.

References

M. Asahara, S. Yasuda, H. Konishi, M. Imada, and K. Maekawa. 2013. BCCWJ-timebank: Temporal and event information annotation on Japanese text. In *the 27th Pacific Asia Conference of Language Information and Computation*, pages 206–214.

N. Asher and A. Lascaridas. 2003. *Logics of Conversation: Studies in Natural Language Processing*. Cambridge University Press.

Jason Baldridge, Nicholas Asher, and Julie Hunter. 2007. Annotation for and robust parsing of discourse structure on unrestricted texts. In *Zeitschrift für Sprachwissenschaft 26.2*, pages 213–239.

Lynn Carlson, John Conroy, Daniel Marcu, Dianne O'Leary, Mary Okurowski, Anthony Taylor, and William Wong. 2001. An empirical study of the relation between abstracts, extracts, and the discourse structure of texts. In *Proceedings of the DUC-2001 Workshop on Text Summarization*.

Greg N Carlson. 1977. *Reference to Kinds in English*. Garland.

Theodore Fernald. 2000. *Predicates and Temporal Arguments*. Oxford University Press.

Gerhard Jäger. 2001. Topic-comment structure and the contrast between stage level and individual level predicates. *Journal of Semantics*, pages 83–126.

Taro Kageyama. 2006. Property description as a voice phenomenon. In Masayoshi Shibatani and Taro Kageyama, editors, *Voice and Grammatical Relations: In Honor of Masayoshi Shibatani*, pages 85–114. John Benjamins.

K. Kaneko and D. Bekki. 2014. Toward a discourse theory for annotating causal relations in Japanese. In *the 28th Pacific Asia Conference of Language Information and Computation*, pages 460–469.

Angelica Kratzer. 1995. Stage-level and individual-level predicates. In Francis Jeffry Pelletier and Greg N Carlson, editors, *The Generic Book*, pages 125–175. The University of Chicago Press.

Manfred Krifka, Francis Pelletier, Greg Carlson, Alice ter Meulen, Godehard Link, and Gennaro Chierchia. 1995. Genericity: An introduction. In Francis Jeffry Pelletier and Greg N Carlson, editors, *The Generic Book*, pages 1–124. The University of Chicago Press.

Kikuo Maekawa, Makoto Yamazaki, Toshinobu Ogiso, Takehiko Maruyama, Hideki Ogura, Wakako Kashino, Hanae Koiso, Masaya Yamaguchi, Makiro Tanaka, and Yasuharu Den. 2014. Balanced corpus of contemporary written Japanese. *Language Resources and Evaluation*, 48(2):345–371.

W. C. Mann and S. Thompson. 1987. Rhetorical structure theory: A theory of text organization. Technical report, University of Southern California, Information Sciences Institute.

Luise McNally. 1998. The stage/individual distinction and (in)alienable possession. In Susan Rothstein, editor, *Events and Grammar*, pages 293–307. Springer Netherlands.

Gary Milsark. 1979. *Existential Sentences in English*. Gerland.

Anne-Lyse Minard, Manuela Speranza, Eneko Agirre, Itziar Aldabe, Marieke van Erp, Bernardo Magnini, and Ruben Urizar German Rigau. 2015. SemEval-2015 task 4: Timeline: Cross-document event ordering. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*.

Yoshiki Ogawa. 2001. The stage/individual distinction and (in) alienable possession. *Language*, pages 1–25.

L. Polanyi, C. Culy, M. Van Den Berg, G. L. Thione, and D. Ahn. 2004. A rule based approach to discourse parsing. In *SIGDIAL Vol. 4*.

R. Prasad, A. Joshi, N. Dinesh, A. Lee, E. Miltsakaki, and B. Webber. 2005. The Penn discourse treebank as a resource for natural language generation. In *the Corpus Linguistics Workshop on Using Corpora for Natural Language Generation*, pages 25–32.

Rashmi Prasad, Bonnie Webber, and Aravind Joshi. 2014. Reflections on the Penn discourse treebank, comparable corpora, and complementary annotation. *Computational Linguistics*, 40(4):921–950.

James Pustejovsky, José M Castano, Robert Ingria, Roser Sauri, Robert J Gaizauskas, Andrea Setzer, Graham Katz, and Dragomir R Radev. 2003a. TimeML: Robust specification of event and temporal expressions in text. *New directions in question answering*, 3:28–34.

James Pustejovsky, David Day, Lisa Ferro, Robert Gaizauskas, Patrick Hanks, Marcia Lazo, Roser Sauri, Andrew See, Andrea Setzer, and Beth Sundheim. 2003b. The TIMEBANK corpus. In *Corpus Linguistics*, pages 647–656.

James Pustejovsky, Marc Verhagen, Xue Nianwen, Robert Gaizauskas, Mark Hepple, Frank Schilder, Graham Katz, Roser Sauri, Estela Saquete, Tommaso Caselli, Nicoletta Calzolari, Kiyong Lee, and Seohyun Im. 2009. TempEval2: Evaluating events, time expressions and temporal relations. In *SemEval-2010 Task Proposal*.

Naushad UzZaman, Hector Llorens, James Allen, Leon Derczynski, Marc Verhagen, and James Pustejovsky. 2012. TempEval-3: Evaluating events, time expressions, and temporal relations. In *SemEval-2013 Task Proposal*.