Building a Japanese Corpus of Temporal-Causal-Discourse Structures Based on SDRT for Extracting Causal Relations

Kimi Kaneko [1] **Daisuke Bekki** [1,2,3]
[1] Ochanomizu University, Tokyo, Japan
[2] National Institute of Informatics, Tokyo, Japan
[3] CREST, Japan Science and Technology Agency, Saitama, Japan
{kaneko.kimi | bekki}@is.ocha.ac.jp

Abstract

This paper proposes a methodology for generating specialized Japanese data sets for the extraction of causal relations, in which temporal, causal and discourse relations at both the fact level and the epistemic level, are annotated. We applied our methodology to a number of text fragments taken from the Balanced Corpus of Contemporary Written Japanese. We evaluated the feasibility of our methodology in terms of agreement and frequencies, and discussed the results of the analysis.

1 Introduction

In recent years, considerable attention has been paid to deep semantic processing. Many studies (Betherd et al., 2008), (Inui et al., 2007), (Inui et al., 2003), (Riaz and Girju, 2013) have been recently conducted on deep semantic processing, and causal relation extraction (CRE) is one of the specific tasks in deep semantic processing. Research on CRE is still developing and there are many obstacles that must be overcome.

Inui *et al.* (2003) acquired cause and effect pairs from text, where the antecedent events were taken as causes and consequent events were taken as effects based on Japanese keywords such as *kara* and *node*. In (1), for example, the antecedent *ame-ga hutta* ('it rained') and the consequent *mizutamari-ga dekita* ('puddles emerged') are acquired as a pair of cause and effect.

(1) Ame-ga hutta-*node*
rain-NOM fall-past-*because*
mizutamari-ga dekita.
puddles-NOM emerge-past
'Because it rained, puddles emerged.'

However, antecedents are not always causes or reasons for consequents in Japanese, as illustrated by the following example.

(2) Zinsinziko-ga
injury.accident-NOM
okita-*kara* densya-ga
happen-past-*because* trains-NOM
tiensita to-iu-wake-dewanai.
delay-past it.is.not.the.case.that
'It is not the case that the trains were delayed because an injury accident happened.'

In example (2), the antecedent *zinsinziko-ga okita* ('an injury accident happened') is not the cause of the consequent *densya-ga tiensita* ('the trains were delayed'). Though in such sentences that contain causal expressions there are no causal relations between antecedents and consequents, in existing studies each sentence containing a causal expression was extracted as knowledge representing cause and effect, such as in (Inui et al., 2003). It is difficult for computers to auto-recognize and exclude such cases.

In this paper, we report on the analysis of necessary information for acquiring more accurate cause-effect knowledge and propose a methodology for creating a Japanese corpus for CRE. First, we introduce previous studies and describe information that should be used to annotate data sets. Next, we describe our methodology based on Segmented Discourse Representation Theory (SDRT) (Asher et al., 2003). Finally, we evaluate the validity of our methodology in terms of agreement and frequency, and analyze the results.

2 Previous Studies

In this section, we introduce previous studies on annotation of temporal, causal and other types of relations and present a linguistic analysis of temporal and causal relations.

Betherd *et al.* (2008) generated English data sets annotated with temporal and causal relations and analyzed interactions between the two types of

Proceedings of the EACL 2014 Workshop on Computational Approaches to Causality in Language, pages 33–39,
Gothenburg, Sweden, April 26, 2014.

relations. In addition, these specialized data sets were evaluated in terms of agreement and accuracy. Relations were classified into two causal categories (CAUSAL, NO-REL) and three temporal categories (BEFORE, AFTER, NO-REL). In regard to the evaluation, Betherd *et al.* pointed out that the classification was coarse-grained, and that reanalysis would have to be performed with more fine-grained relations.

Inui *et al.* (2005) characterized causal expressions in Japanese text and built Japanese corpus with tagged causal relations. However, usages such as that illustrated in (2) and interactions between temporal relations and causal relations were not analyzed.

Tamura (2012) linguistically analyzed temporal and causal relations and pointed out that in reason/purpose constructions in Japanese, the event time indicated by the tense sometimes contradicts the actual event time, and that the information necessary to recognize the order between events lies in the choice of the *fact* and the *epistemic* levels (we will come back to these notions in the section 3.4), and the explicit or implicit meaning of a sentence in the causal expressions in Japanese. Furthermore, some causal expressions in Japanese are free from the absolute and relative tense systems, and both the past and non-past forms can be freely used in main and subordinate clauses (Chin, 1984) (an example is given in the next section). In other words, temporal relations are not always resolved earlier than causal relations, and therefore we should resolve temporal relations and causal relations simultaneously.

Asher *et al.* (2003) proposed SDRT in order to account for cases where discourse relations affect the truth condition of sentences. Because temporal relations constrain causal relations, the explicit or implicit meaning of a sentences and the epistemic level information affects preceding and following temporal relations in causal expressions in Japanese, recognition also affects causal relations. Therefore, the annotation of both causal relations and discourse relations in corpora is expected to be useful for CRE. Moreover, which characteristics (such as tense, actual event time, time when the event is recognized, meaning and structure of the sentence and causal relations) will serve as input and which of them will serve as output depends on the time and place. Therefore, we should also take into account discourse relations together with temporal and causal relations. We can create specialized data sets for evaluating these types of information together by annotating text with discourse, temporal and causal relations.

However, discourse relations of SDRT are not distributed into discourse relations and temporal relations, and as a result the classification of labels becomes unnecessarily complex. Therefore, it is necessary to rearrange discourse relations as in the following example.

(3) Inu-wa niwa-o kakemawatta.
dog-NOM garden-ACC run-past
Neko-wa $kotatu$-de
cat-NOM $kotatsu$.heater-LOC
marukunatte-ita.
be.curled.up-past
'The dog ran in the garden. The cat was curled up in the *kotatsu* heater.'

This pair of sentences is an antithesis, so we annotate it with the "Contrast" label in SDRT. On the other hand, the situation described in the first sentence overlaps with that of the second sentence, so we annotate this pair of sentences with the "Background" label as well. Though there are many cases in which we can annotate a sentence with discourse relations in this way, dividing temporal relations from discourse relations as in this study allows us to avoid overlapping discourse relations.

This study was performed with the aim to rearrange SDRT according to discourse relations, temporal relations and causal relations separately, and we generated specialized data sets according to our methodology. In addition, occasionally it is necessary to handle the actual event time and the time when the event was recognized individually. An example is given below.

(4) Asu tesuto-ga
tomorrow exam-NOM
aru-*node*, kyoo-wa
take.place-nonpast-*because*, today-TOP
benkyoo-suru-koto-ni sita.
to.study-DAT decide-past
'Because there will be an exam tomorrow, I decided to study today.

Before we evaluate the consequent *kyoo-wa benkyoo-suru-koto-ni sita* ('I decided to study today'), we should recognize the fact of the antecedent *Asu tesuto-ga aru* ('there will be an exam tomorrow'). Whether we deal with the actual

Label	Description
Precedence(A,B)	End time (A) $<$ start time (B) In other words, event A temporally precedes event B.
Overlap(A,B)	Start time (A) $<$ end time (B) $\leq$ end time (B) $<$ end time (A), In other words, event A temporally overlaps with event B.
Subsumption(A,B)	Start time (A) $\leq$ end time (B) & End time (A) $\leq$ end time (B), In other words, event A temporally subsumes event B.

Table 1: Temporal relations list

Level	Description
Cause(A,B)	The event in A and the event in B are in a causal relation.

Table 2: Causal relation

event time or the time when the event was recognized depends on the circumstances. Therefore, we decided to annotate text at the fact and epistemic levels in parallel to account for such a distinction.

3 Methodology

We extended and refined SDRT and developed our own methodology for annotating main and subordinate clauses, phrases located between main and subordinate clauses (e.g., continuative conjuncts in Japanese), two consecutive sentences and two adjoining nodes with a discourse relation. We also defined our own method for annotating propositions with causal and temporal relations. The result of tagging example (5a) is shown in (5b).

(5) a. Kaze-ga huita. Harigami-ga
wind-NOM blow-past poster-NOM
hagare, tonda.
come.off-past flow-past
'The wind blew. A poster came off and flew away.'

b. **[Precedence**(π1,π3)**,Explanation**(π1,π3)**,**
Cause(π1,π3)**]**,
[Precedence(π2,π4)**, Explanation**(π2,π4)**,**
Cause(π2,π4)**]**
$_{\pi 2 \pi 1}$Kaze-ga huita.
$_{\pi 4 \pi 3}$Harigami-ga hagare, tonda.

The remainder of this section is structured as follows. Sections 3.1 and 3.2 deal with temporal and causal relations, respectively. Section 3.3 covers discourse relations, and Section 3.4 describes the fact level and the epistemic level.

3.1 Temporal Relations

We consider the following three temporal relations (Table 1). We assume that they represent the relations between two events in propositions and indicate a start time and an end time. In addition, we also assume that (start time of e) $\leq$ (end time of e) for all events. Based on this, the temporal placement of each two events is limited to the three relations in Table 1.

In this regard, Japanese non-past predicates occasionally express habitually repeating events, which have to be distinguished from events occurring later than the reference point. In this paper, in annotating the scope of the repetition, habitually repeating events are described as in the following example.

(6) a. Taiin-go, {kouen-o
After.retirement park-ACC
hasiru}$_{repeat}$ yoo-ni-site-iru.
to.run have.a.custom
'After retiring, I have a custom to {run in the park}$_{repeat}$.'

b. {supootu-inryo-o nonda-ato,
Sports.drink-ACC drink-past-after
kouen-o hasiru}$_{repeat}$
park-ACC run
yoo-ni-site-iru.
have.a.custom
'I have a custom that {I run in the park after having a sports drink}$_{repeat}$.'

3.2 Causal Relations

We tag pairs of clauses with the following relation (Table 2) only if there is a causal relation between events in the proposition. By annotating text with discourse relations, a fact and epistemic level and temporal relations, we can describe the presence

Label	Description
Alternation(A,B)	"A or B", where the pair of A and B corresponds to logical disjunction ($\vee$).
Consequence(A,B)	"If A then B", where the pair of A and B corresponds to logical implication ($\rightarrow$).
Elaboration(A,B)	B explains A in detail in the discourse relation. B of the event is part of A of the event.
Narration(A,B)	A and B are in the same situation, and the pair of A and B corresponds to logical conjunction ($\wedge$).
Explanation(A,B)	The discourse relation indicates A as a cause and B as an effect.
Contrast(A,B)	"A but B", where A and B are paradoxical.
Commentary(A,B)	The content of A is summarized or complemented in B.

Table 3: Discourse relations list

SDRT	Our methodology	Rules
Alternation(A,B)	Alternation(A,B)	NA
Consequence(A,B)	Consequence(A,B)	NA
Elaboration(A,B)	Elaboration(A,B)	$\forall$ A,B (Elaboration(A,B) $\rightarrow$ Subsumption (A,B))
Narration(A,B)	Precedence(A,B) $\wedge$ Narration(A,B)	NA
Background(A,B)	Subsumption(A,B) $\wedge$ Narration(A,B)	NA
Result(A,B) **Explanation(A,B)**	Explanation(A,B) Cause(A,B)	$\forall$ A,B (Cause(A,B) $\rightarrow$ Temp_rel(A,B)) [1]
Contrast(A,B)	Contrast(A,B)	NA
Commentary(A,B)	Commentary(A,B)	NA

Table 4: Correspondence between SDRT and our methodology

of causation in finer detail than (Betherd et al., 2008).

3.3 Discourse Relations

We consider the following discourse relations based on SDRT (Table 3). There are also relations that impose limitations on temporal and causal relations (Table 4). The way temporal, causal and discourse relations affect each other is described below together with their correspondence to the relations in SDRT. **Bold-faced** entries represent relations integrated in SDRT in our study.
Such limitations on temporal relations provides information for making a decision in terms of temporal order and cause/effect in the "de-tensed" sentence structure [2] (Chin, 1984) in Japanese. An example is given below.

(7) Kinoo anna-ni taberu-$\underline{kara}$,
yesterday that.much eat-past-$\underline{because}$
kyoo onaka-ga itaku
today stomach-NOM ache-cont
natta-nda.
become-$noda$
'Because you ate that much yesterday, you have a stomachache today.'

(7) **[Precedence(π1,π3),Explanation(π1,π3), Cause(π1,π3)],**
[Precedence(π2,π4),Explanation(π2,π4), Cause(π2,π4)]
$\pi 2 \pi 1$Kinoo anna-ni taberu-$kara$,
$\pi 4 \pi 3$kyoo onaka-ga itaku natta-nda.

This is a sentence where the subordinate clause is in non-past tense and the main clause is in past tense. Then, we may mistakenly interpret the event in the subordinate clause as occurring after the event of the main clause. However, we can determine that in fact it occurred *before* the event in the main clause based on the rule imposed by the "Cause" relation.

3.4 Fact Level and Epistemic Level

A fact level proposition refers to an event and its states, while an epistemic level proposition refers to speaker's *recognizing* event of a described event. In Japanese, the latter form is often marked by the suffix *noda* that attaches to all kinds of predicates (which may also be omitted). Both overt and covert *noda* introduce embedded structures, and we annotate them in such a way that a fact level proposition is embedded in an epistemic level proposition.

Semantically, the most notable difference between the two levels is that the tense in the former

[2] Temp_rel(A,B) $\equiv$
Precedence(A,B)$\vee$ Overlap(A,B) $\vee$ Subsumption(A,B)

[3] According to (Chin, 1984), "de-tensed" is a relation whereby the phrase has lost the meaning contributed by tense, namely, the logical aspect of the semantic relation between an antecedent and a consequent has eliminated the aspect temporal relation between them.

represents the time that an event takes place, while the tense in the latter represents the time that the speaker *recognizes* the event.

This distinction between the two types of propositions is carried over to the distinction between the fact level and the epistemic level causal relations. We annotate the former by the tag "Cause" and the latter by the tag "Explanation".

In Japanese, a causal marker such as *node* (a continuation form of *noda*) and *kara* are both used in the fact level and the epistemic level. The fact level causality is a causal relation between the two events, while the epistemic level causality is a causal relation between the two *recognizing* events of the two events mentioned. Therefore, in the causal construction, it happens that the precedence relations between the subordinate and the matrix clauses in the fact level and the epistemic level do not coincide, as in the following example.

(8) Kesa nani-mo
this.morning nothing-NOM
hoodoo-sare-nakatta-$\underline{node}$,
report-passive-NEG.past-$\underline{because}$,
kinoo-wa mebosii ziken-wa
yesterday-TOP notable events-NOM
nakatta-noda.
be-NEG-$noda$
'Because nothing was reported this morning, there were no notable event yesterday.'

[Precedence($\pi3$,$\pi1$)**,Explanation**($\pi3$,$\pi1$)**,**
Cause($\pi3$,$\pi1$)**]**,
[Precedence($\pi2$,$\pi4$)**, Explanation**($\pi2$,$\pi4$)**,**
Cause($\pi2$,$\pi4$)**]**
$_{\pi2\pi1}$Kesa nani-mo hoodoo-sare-nakatta-$node$, $_{\pi4\pi3}$kinoo-wa mebosii ziken-wa nakatta-noda.

The temporal relation at the fact level is that $\pi3$ precedes $\pi1$. By contrast, that at the epistemic level is that $\pi2$ precedes $\pi4$. By describing the relation between $\pi1$ and $\pi3$ and that between $\pi2$ and $\pi4$ separately, we can reproduce the relationship at both levels.

3.5 Merits

We defined our methodology for annotating text fragments at both the fact and epistemic levels in parallel with temporal, causal and discourse relations. Therefore, we can generate specialized data sets that enable estimating the causality in the fact and epistemic levels by various cues (such as known causal relations, truth condition, conjunctions and temporal relations between sentences or clauses).

In addition, we can say that causal expressions without causation are not in a causal relation (and vice versa) by annotating text with both discourse and causal relations.

4 Results

We applied our methodology to 66 sentences from the Balanced Corpus of Contemporary Written Japanese (BCCWJ) (Maekawa, 2008). The sentences were decomposed by one annotator, and labels were assigned to the decomposed segments by two annotators. During labeling, we used the labels presented in Section 3. Our methodology was developed based on 96 segments (38 sentences), and by using the other 100 segments (28 sentences), we evaluated the inter-annotator agreement as well as the frequencies of decomposition and times of annotation. The agreement for 196 segments generated from 28 sentences amounted to 0.68 and was computed as follows (the kappa coefficient for them amounted to 0.79).

$$Agreement = Agreed\ labels / Total\ labels$$

Analyzing more segments in actual text and improving our methodology can lead to further improvement in terms of agreement.

Table 5 shows the distribution of labels into segments in our study.

label	**segments**		
	Total	fact	epistemic
Precedence	25	14	11
Overlap	7	4	3
Subsumption	61	29	32
total	**94**	**47**	**47**
Cause	14	8	6
total	**14**	**8**	**6**
Alternation	–	–	–
Consequence	6	3	3
Elaboration	4	2	2
Narration	66	33	33
Explanation	14	7	7
Contrast	2	1	1
Commentary	**94**	**47**	**47**

Table 5: Distribution of labels in segments in our study

We can see from Table 5 that "Narration" was the most frequent one, while "Alternation" never appeared. As s result, we can assume that frequent relations will be separated from non-frequent relations. So far, all the relations are either frequent or non-frequent. We should re-analyze the data with more samples again.

When the methodology was applied to 28 sentences, a total of 100 and an average of 3.57 segments were derived. This is the number of segments at both the fact and epistemic levels. Without dividing the fact and epistemic levels, an average of 1.79 segments were derived.

On average, 11 segments per hour were tagged in our study. Although we should evaluate the validity after having computed the average decomposition times, it is assumed that our methodology is valid when focusing only on labeling.

5 Discussion

We analyzed errors in this annotation exercise. The annotators often found difficulties in judging temporal relations in the following two cases: (1) the case where it was difficult to determine the scope of the segments pairing and (2) the case where formalization of lexical meaning is difficult.

In regard to the first case, how to divide segments sometimes affects temporal relations. In the following example, consider the temporal relation between the first and the second sentences.

(9) Marason-ni syutuzyoo-sita.
marathon-DAT participate-past.
sonohi-wa 6zi-ni kisyoo-si,
that.day-TOP 6:00-at get.up-past,
10zi-ni totyoo-kara
10:00-at Metropolitan.Government-from
syuppatu-site, 12zi-ni
leave-past, 12:00-at
kansoo-sita.
finish.running-past.
'I participated in marathon. I got up at 6:00 on that day and left the Metropolitan Government at 10:00 and finished running at 12:00.'

When we focus on the first segment of the second sentnce ('*I got up at 6:00*'), its relation to the first sentence appears to be "Precedence". However, if we consider the second and the third segments as the same segment, their relation to the first sentence appears to be "Subsumption".

Therefore, we should establish clear criteria for the segmentation. Although we currently adopts a criterion that we chose smaller segment in unclear cases, there still remain 9 unclear cases (temporal:5, discourse:4).

One of the reason why Kappa coefficient marks relatively high score is that we only compare the labels and ignore the difference in the segmentations. Criteria for deciding the segment scope in paring segments will improve our methodology.

The second case is exemplified by the temporal relation between the subordinate clause and the main clause in the following sentence.

(10) Migawari-no tomo-o
scapegoat-GEN friend-ACC
sukuu-*tame-ni* hasiru-noda.
to.save run-$noda$.
'I run to save my friend who is my scapegoat.'

If we consider that the *saving* event only spans over the very moment of *saving*, the relation between the clauses appears to be "Precedence". However, if we consider that *running* event is a part of the *saving* event, the relation between the clauses is "Subsumption".

Thus, judging lexical meaning with respect to when events start and end involves some difficulties and they yield delicate cases in judging temporal relations.

These problems are mutually related, and the first problem arises when the components of a lexical meaning are displayed explicitly in the sentence, and the second problem arises when they are implicit.

6 Conclusions

We analyzed and proposed our methodology based on SDRT for building a more precise Japanese corpus for CRE. In addition, we annotated 196 segments (66 sentences) in BCCWJ with temporal relations, discourse relations, causal relations and fact level and epistemic level propositions and evaluated the annotations of 100 segments (28 sentences) in terms of agreement, frequencies and times for decompositions. We reported and analyzed the result and discussed problems of our methodology.

The discrepancies of decomposition patterns were not yet empirically compared in the present study and will be investigated in future work.

References

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

Bethard S., Corvey W. and Kilingenstein S. 2008. *Building a Corpus of Temporal Causal Structure*. LREC 2008, Marrakech, Morocco.

Chin M. 1984. *Tense of the predicates for clauses of compound statement binded by conjunctive particle -"Suru-Ga" and "Shita-Ga", "Suru-Node" and "Shita-Node" etc.-*. Language Teaching Research Article.

Inui T., Inui K. and Matsumoto Y. 2005. *Acquiring Causal Knowledge from Text Using the Connective Marker Tame*. ACM Transactions on Asian Language Information Processing (ACM-TALIP), Vol.4, Issue 4, Special Issue on Recent Advances in Information Processing and Access for Japanese, 435–474.

Inui T., Inui K. and Matsumoto Y. 2003. *What Kinds and Amounts of Causal Knowledge Can Be Aquired from Text by Using Connective Markers as Clues*. The 6th International Conference on Discovery Science (DS-2003), 180–193.

Inui T., Takamura H. and Okumura M. 2007. *Latent Variable Models for Causal Knowledge Acquisition. Alexander Gelbukh(Ed.), Computational Linguistics and Intelligent Text Processing, Lecture Notes in Computer Science*, 4393:85–96.

Maekawa K. 2008. *Balanced Corpus of Contemporary Written Japanese*. In Proceedings of the 6th Workshop on Asian Language Resources (ALR), 101—102.

Riaz M. and Girju R. 2013. *Toward a Better Understanding of Causality between Verbal Events:Extraction and Analysis of the Causal Power of Verb-Verb Associations*. In Proceedings of the SIGDIAL 2013 Conference, Metz, France 21–30.

Tamura S. 2012. *Causal relations and epistemic perspectives: Studies on Japanese causal and purposive constructions*. Doctoral thesis, Kyoto University.