Problems for successful bunsetsu based parsing and some solutions

Alastair BUTLER*† Zhen ZHOU‡ Kei YOSHIMOTO†‡
*PRESTO, Japan Science and Technology Agency
†Center for the Advancement of Higher Education, Tohoku University
‡Graduate School of International Cultural Studies, Tohoku University

ajb129@hotmail.com

Abstract
Parsing is essential to reveal internal linguistic structure for advanced NLP applications. For Japanese tremendous effort has gone into creating corpora, heuristics, parsing algorithms and statistical based learning methods to support bunsetsu based dependency parsing. Achievements are hugely impressive, both in the scale of corpora and success scores of parsers. In this paper we address whether bunsetsu based parsing is sufficient to feed further processing, with a focus on deriving meaning representations with explicit scoped operations of quantification and their bindings. We detail areas in which even successful bunsetsu parsing simply fails to provide essential structural information, notably failing to distinguish cases of subordination / embedding from coordination and vice versa. We do however show that a large amount of missing information can be recovered when case frame information is also available. This brings us steps closer to confirming that sufficiently rich parsing information can be obtained for semantic processing via the bunsetsu parsing route, but problems remain concerning the determination of scope for non-final scopal operations like negation.

1 Introduction

Introduced by Hashimoto (1934) a *bunsetsu* is a phrasal unit consisting of one or more adjoining content words (noun, verb, adjective, etc.) and zero or more functional words (postposition, auxiliary verb, etc.). A bunsetsu dependency analysis involves segmenting the sentence into bunsetsu and establishing modifier (dependence on) relations between the bunsetsu. Such analysis reveals information about the internal structure of sentences. The question addressed in this paper is whether the revealed internal structure is suitable to feed building meaning representations for advanced NLP applications. When information is found to be lacking, we aim to propose ways to supplement the bunsetsu analysis.

The paper is structured as follows. Section 2 touches on the challenge of segmenting bunsetsu. Section 3 shows an inability to distinguish embedded clauses from relative clauses and vice versa, but offers a potential solution relying on the presence of case information. The problem of section 3, and potential solution, carry over to section 4, arising whenever a sentence is complex. In section 5 we consider sentence final particles, arguing they should be analysed as forming the last bunsetsu of a sentence in order to scope over the sentence. Section 6 observes bunsetsu analysis offers no basis to determine the scope of non-final scopal operators like negation. Section 7 provides a summary.

2 Segmenting Bunsetsu

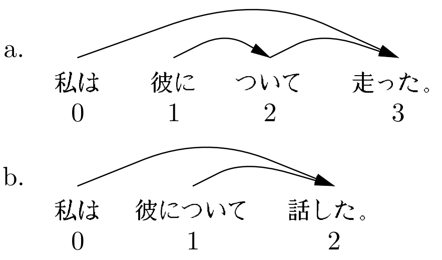
A big challenge for bunsetsu analysis is deciding the segmentation. This is particularly so for combinations of morphemes that may have either a literal interpretation arising from the combination of elements or a compound functional role. To see this, consider について in (1).

(1) a. 私は彼について走った。
I ran following him.

b. 私は彼について話した。
I talked about him.

In (1a) particle に has the case-marking function of 'with' modifying the verb ついて 'keep close contact' to produce the literal content meaning of 'follow'. By contrast, in (1b) について has a case-marking function as a single unit similar to 'about'. Bunsetsu analysis needs to be sensitive to such distinctions to give, for example, the analysis of (2a) for (1a) and (2b) for (1b).

(2)



The analyses of (2) offer structural information to build the meaning representations of (3).

(3) a. $\exists e_1 e_2 x(\text{彼}(x) \wedge \text{つく}(e_1,\text{私}) \wedge \text{に}(e_1) = x \wedge \text{走る}(e_2,\text{私}))$

b. $\exists e_1 x(\text{彼}(x) \wedge \text{話す}(e_1,\text{私}) \wedge \text{について}(e_1) = x)$

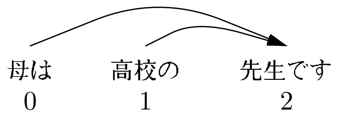
This assumes a Davidsonian theory (Davidson, 1967) where verbs are predicates with minimally an implicit event argument. The verbs of (3) also have subject arguments. Moreover events are existentially quantified over and may be further constrained. In (3a) e_1 is an event occurring with (に) some value that is restricted to be 彼 'him'. In (3b) e_1 is coded to be an event that occurs about (について) some value restricted to be 彼 'him'.

Next consider the copula sentence of (4).

(4) 母は高校の先生です。
My mother is a high school teacher.

The copula is typically taken to be a functional word allowing only the analysis of (5).

(5)



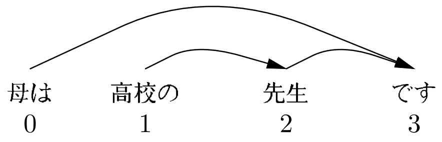
But with (5) it is difficult to read off the correct semantics. That is, while the meaning representation (6) is readily derivable, the appropriate meaning representation of (7) is not.

(6) $\exists x(\text{母}(x) \wedge \exists e_1 y(\text{高校}(y) \wedge \text{先生\_です}(e_1,x) \wedge \text{の}(e_1) = y))$

(7) $\exists x(\text{母}(x) \wedge \exists e_1 yz(\text{高校}(y) \wedge \text{の\_先生}(z,y) \wedge \text{です}(e_1,x,z)))$

To derive the meaning representation of (7) we need an analysis along the lines of (8) where the copula is taken to be a content word that is able to form a bunsetsu by itself.

(8)



3 Relative Clause vs. Embedded Clause

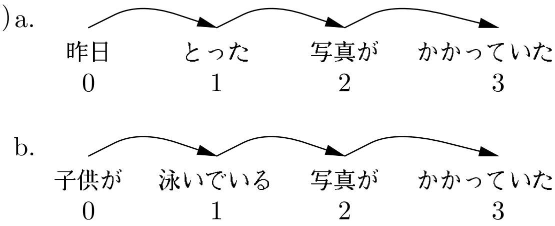
In (9a) 昨日とった 'we took yesterday' is a relative clause that modifies 写真 'picture'. By contrast in (9b) 子供が泳いでいる 'a swimming child' is an embedded clause, and is the content of 写真 'the picture'.

(9) a. 昨日とった写真がかかっていた。
'The picture that we took yesterday was hung.'

b. 子供が泳いでいる写真がかかっていた。
'The picture of a swimming child was hung.'

Bunsetsu dependency analyses for the sentences of (9) are given in (10).

(10)



The role of dependencies between 0 and 1 and between 2 and 3 are readily determined, either from case-marker が or from the expected function of a bunsetsu that is a nominal denoting a time. That is, such dependencies stem from noun phrases contributing arguments. However the dependencies between 1 and 2 are not obvious from either the structural information of the bunsetsu analyses or from the bunsetsu content: for both sentences bunsetsu 1 is a predicate bunsetsu and bunsetsu 2 is a nominal bunsetsu. This is unfortunate since detecting the dependency contribution has a dramatic consequence for the resulting meaning representation. Thus the relative clause dependency of (9a) leads to meaning representation (11a), while the embedded clause dependency of (9b) should produce (11b).

(11) a. $\exists e_1 x(\text{写真}(x) \wedge \exists e_2(\text{とる}(e_2,\_,x) \wedge \text{時間}(e_2) \sqsubseteq \text{昨日}) \wedge \text{かかる}(e_1,x))$

b. $\exists e_1 x(\text{写真}(x,\exists e_2 y(\text{子供}(y) \wedge \text{泳ぐ}(e_2,y))) \wedge \text{かかる}(e_1,x))$

To distinguish between a relative clause dependency and an embedded clause dependency we must look for information to supplement the bunsetsu dependency analyses. In this regard case information, if available, would provide relevant extra information. For example, case information might allow us to conclude that 昨日とった 'we took yesterday' is a clause requiring an object binding to trigger a relative clause analysis, while 子供が泳いでいる 'a swimming child' is a saturated clause thereby triggering an embedded clause analysis.

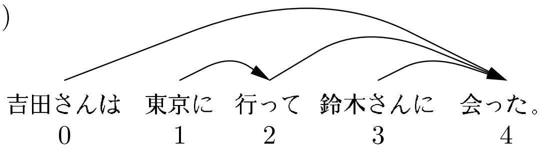
4 Complex Sentences

A complex sentence contains more than one clause, raising the issue of how the clauses combine to make up the sentence. Clauses may combine with coordinate conjunctions such as が 'but' or with the て-forms of verbs, adjectives or the copula meaning '∼ and', as in (12).

(12) 吉田さんは東京に行って鈴木さんに会った。
Mr. Yoshida went to Tokyo and met Mr. Suzuki.

A dependency analysis for (12) is given by (13), from which we might hope to build (14), which combines with conjunction the contributions of the clauses that make up (12).

(13)

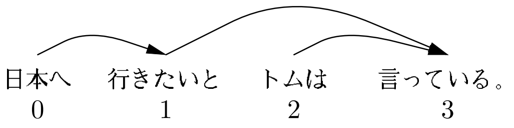


(14) $\exists e_1 e_2$(行く(e_1,吉田) $\wedge$ ニ(e_1) $=$ 東京 $\wedge$
会う(e_2,吉田) $\wedge$ ニ(e_2) $=$ 鈴木)

A different way to combine clauses of a complex sentence is illustrated by (15), which we might expect to receive the bunsetsu dependency analysis of (16).

(15) 日本へ行きたいとトムは言っている。
Tom says that he wants to go to Japan.

(16)



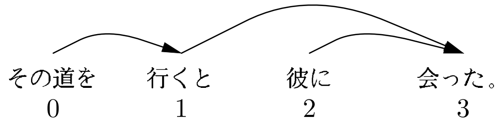
For (15) we want to build meaning representation (17), but we arrive at the same problem we had in distinguishing embedded clauses from relative clause and vice versa, that is we have a dependency but we also need to spell out the role of the dependency sufficiently to determine whether this leads to an instance of embedding, as in (17), or of coordination, as in (14).

(17) $\exists e_1$言う(e_1,トム,
$\exists e_2$(行く(e_2) $\wedge$ へ(e_2) $=$ 日本))

Note we cannot rely on the presence of particle と in (15) to conclude the presence of an embedded clause, since と also has a subordinate conjunction function, as (18) demonstrates. From the dependency analysis of (19) we need to be able to produce (20).

(18) その道を行くと彼に会った。
'As I went along the road, I met him.'

(19)



(20) $\exists e_1 e_2 xy$(その道(x) $\wedge$ 彼(y) $\wedge$
と(行く(e_1,_,x),会う(e_2) $\wedge$ ニ(e_2) $=$ y))

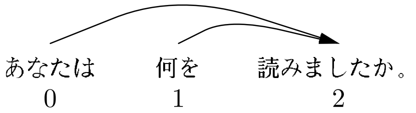
As in the previous section we find that a bunsetsu dependency analysis fails to offer sufficient information and so we must look for supplementary information. Again relevant information might come from case information if available, with for example the case frame of 言う 'says' offering a signal to trigger the embedding seen with (17).

5 Sentence Final Particles

Requiring bunsetsu to contain a content word forces a sentence final particle, such as the question particle か, to be part of the last bunsetsu. Thus (21) is analysed as (22).

(21) あなたは何を読みましたか。
What did you read?

(22)



With (22) か as a scope taking operator can be placed no higher than the verb, notably falling under the scope of the questioned argument, as in (23).

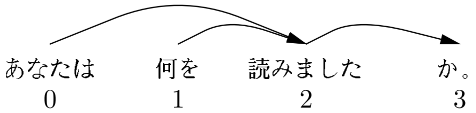
(23) $\exists x$(あなた(x) $\wedge$? y(何(y) $\wedge$
$\exists e_1$(か(読む(e_1,x,y)))))

Representation (24) is more suitable, with か taking scope above the questioned argument.

(24) $\exists x$(あなた(x) $\wedge$ か(? y(何(y) $\wedge$
$\exists e_1$(読む(e_1,x,y)))))

One way to achieve the form of (24) would be to adopt the dependency analysis of (25), with the sentence final particle as a distinct 'bunsetsu' that is the root of the sentence.

(25)



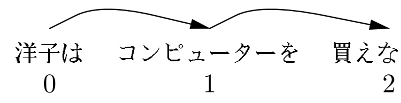
Note that in (24) the topic argument scopes outside か. This is derivable from (25) with the assumption that topic marking serves to promote an argument to discourse level scope irrespective of where it occurs in the structure of the sentence.

6 Negation

Following the analysis of sentence final particles in the previous section we might aim to treat other elements that form scopal operations in a meaning representation as separate ‘bunsetsu’. For example ない ‘not’ is typically analysed as being part of a larger bunsetsu that contains a predicate as the content word, so it would be standard to analyse (26) as in (27).

(26) 洋子はコンピューターを買えない。
Yoko is unable to buy a computer.

(27)



With (27) as a basis for building a meaning representation the scope of negation is restricted to scope only over the verb, to for example derive the meaning representation of (28).

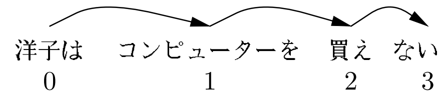
(28) $\exists e_1 x$(コンピューター(x) $\wedge$
ない(買える(e_1,洋子,x)))

For (28) to be true there should be some computer and some event such that the event is not Yoko buying the computer. By contrast (26) will be true if there are no computers. This is captured by (29) where negation scopes over the quantification of computer and event.

(29) ない($\exists e_1 x$(コンピューター(x) $\wedge$
買える(e_1,洋子,x)))

To derive (29) we need an analysis along the lines of (30), with ない as a distinct ‘bunsetsu’ that is also the root of the sentence.

(30)

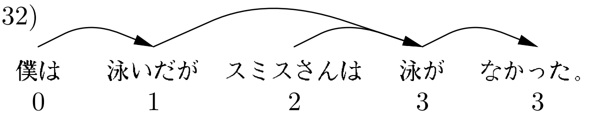


However negation is not a sentence final particle and we find instances where scoping widest is inappropriate:

(31) 僕は泳いだがスミスさんは泳がなかった。
I swam but Mr. Smith didn’t.

We do not want an analysis of (31) to be along the lines of (32) since we would be unable to avoid deriving the representation of (33), in which the first conjunct erroneously falls under the scope of negation.

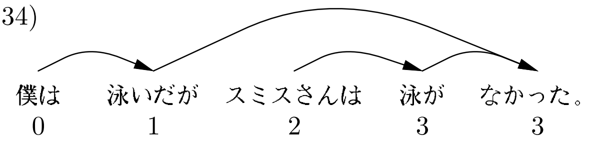
(32)



(33) $\exists x$(僕(x) $\wedge$ ない($\exists e_1 e_2$が(泳ぐ(e_1,x),
泳ぐ(e_2,スミス))))

Rather (34) is the appropriate analysis to allow building the meaning representation of (35).

(34)



(35) $\exists x$(僕(x) $\wedge$ $\exists e_1 e_2$が(泳ぐ(e_1,x),
ない(泳ぐ(e_2,スミス))))

The problem this raises is how should we find the information to appropriately fix the scope of negation.

7 Summary

To sum up we have considered a number of constructions in which a bunsetsu analysis fails to offer information to support constructing appropriate meaning representations deterministically. We looked into providing supplementary information and found case information to be extremely relevant, if available, for deciding whether a bunsetsu dependency contributes subordinate / embedded content or content for coordination. However Sasano, Kawahara and Kurohashi (2010) show recovering case information for sentences of Japanese to be an extremely difficult problem. We also argued how it is necessary to treat elements that form scopal operations in a meaning representation, such as sentence final particles and negation, as distinct ‘bunsetsu’. We left as a problem for further research finding criteria appropriate to fix the scope of non-final scope operations like negation.

This research has been supported by the JST PRESTO program (Synthesis of Knowledge for Information Oriented Society).

References

Davidson, Donald. 1967. The logical form of action sentences. In N. Rescher, ed., *The Logic of Decision and Action*. Pittsburgh: University of Pittsburgh Press. Reprinted in: D. Davidson, 1980. *Essays on Actions and Events*. Claredon Press, Oxford, pages 105–122.

Hashimoto, Shinkichi. 1934. *Essentials of Japanese Grammar (Kokugoho Yousetsu)*. Iwanami. (In Japanese).

Sasano, Ryohei, Daisuke Kawahara, and Sadao Kurohashi. 2010. The effect of corpus size on case frame acquisition for predicate-argument structure analysis. In *IEICE TRANSACTIONS on Information and Systems*, vol. E93-D, no. 6, pages 1361–1368.