自 然 言 語 処 理 117－1
(1997. 1. 20)

日本語の関係節における主辞の省略の解析

Timothy Baldwin, 田中穂積, 徳永健伸
東京工業大学大学院情報理工学研究科

概要

本論文では，関係節の主辞が関係節内で省略されている格要素になるかどうかを認識し，省略として認識された場合には，そのギャップとなっている空格も同定するアルゴリズムを提案する．アルゴリズムは主に動詞の他動詞性と統語的・語彙的フィルターに基づいているが，意味的な情報に関するヒューリスティクスも使用する．129 文章を解析した結果，空格の導出も含めて 90%の精度を得た．

Analysis of Head Gapping in Japanese Relative Clauses

Timothy Baldwin, Hozumi Tanaka, Takenobu Tokunaga

Tokyo Institute of Technology
Dept of Information Engineering

Abstract

This paper describes ongoing research into the determination of the gap associated with the head of a relative clause, and describes a simple algorithm to model this determination process. The method is based on verb transitivity, and on-line syntactic and lexical filtering, but draws partly from semantic analysis. The algorithm determines whether the head of a given relative clause has been gapped from the clause body, and for gapping clauses, returns the case slot associated with the gap. The overall accuracy of the algorithm on a set of 129 randomly selected sentences was around 90%.

0 Introduction

As with most languages, Japanese involves the extensive usage of relative clause constructions, in a number of distinct discourse and semantic roles. Whilst various attempts have been made to classify the syntactic and semantic scope of their usage (see (Tsujimura 1996; Kitagawa 1982; Sato 1989) for representative works), no computational techniques exist to model this classification process. Given that approximately 50% of sentences contain one or more relative clauses, and that there are approximately two relative clauses for every three sentences in an overall text,[1] the accurate analysis of relative clauses would seem to be a vital element of any overall system.

This paper presents an initial attempt to use limited linguistic information to classify relative clauses, based on the parameter of 'head gapping'. Head gapping is defined by the relationship between the clause head and relative clause itself, and whether the head is gapped from within the relative clause or not. Gapped relative clauses are then further categorised according to the case slot identity of the gap.

First, we give our definition of Japanese relative clauses, and then detail our interpretation of gapping within the context of relative clauses. This is followed by details of the algorithm, and particular lexical and syntactic factors relied on in the algorithm. Finally, we present an evaluation of our algorithm on a number of separate data sets, and a discussion of the future direction of our research.

1 The structure of Japanese relative clauses

The general syntactic structure of Japanese relative clause constructions is given in figure 1 below. The noun phrase (NP) clause head is modified by a verb phrase (VP) clause body in comprising the overall noun phrase.

[1] The ratios given are based on an analysis of the EDR corpus (EDR 1995). Of the total of 201340 sentences, 99673 contained one or more relative clauses, and the total number of relative clauses was 133655.

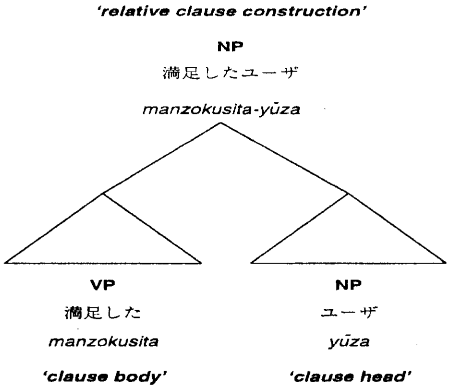

Figure 1: The general form of Japanese relative clauses

Our current research is aimed at verb-based relative clauses, but relative clauses incorporating adjectives and adjectival verbs are theoretically included under the generic category of VP, with the proviso that a morphological connective is required between the clause body and head in the case of adjectival verbs.

Semantically, the clause head and clause body stem verb cannot include clause-level grammatical constructions or grammatical markers which cannot exist as an independent NP. This leads to the preclusion of the following types of VP NP contructions from our definition of relative clauses:

a. NPs such as *noti* (後 'after'), *tame* (ため 'for') and *bāi* (場合 'case/circumstance'). These act as discourse/clause-level markers and unambiguously collocate with a clause body, deictic marker or noun specifier.

b. 'Relational' verb stems, as taken in the Hallidayean sense (see Halliday (1994:119-38)). This includes constructs of the type *to-iu* (という 'called'), *ni-kansuru* (に関する 'concerning') and *ni-taisuru* (に対する 'against/regarding'). Again, these types of verbs cannot exist independently as a main verb.

2 Definition of head gapping

Head gapping is defined as the process whereby the head of the relative clause has been moved from a case slot in the relative clause body, and that case slot is syntactically ellipted. Note that the particular type of the case slot will be defined by the case frame of the root verb, but that any complement argument can be gapped.[2]

Sentence (1) is an example of a relative clause which displays head-gapping, with the associated gap being the nominative case slot. Sentence (2), on the other hand, is an example of a non-gapping clause.[3]

(1)	ϕ_i	満足した	ユーザ$_i$
		manzoku-sita	*yūza*
	SUBJ	to be satisfied-PERF	user

'a satisfied user' (*lit.* a user who is satisfied)

(2)	ϕ_i	個人	の	データ	の	流れ	を	制限する	権利$_j$
		kozin	*no*	*dēta*	*no*	*nagare*	*wo*	*seigen-suru*	*kenri*
	SUBJ	individual	GEN	data	of	flow	ACC	restrict-PRES	right

'the right to restrict the flow of personal data'

[2] The actual process to determine which case slots can be gapped is left as a matter for future research, but it seems uncontentious to state that complements can potentially be gapped in all circumstances.

[3] The following abbreviations are used throughout this paper: NOM = nominative, ACC = accusative, SUBJ = subject, GEN = genitive, TEMP = temporal, PRES = present, PROG = progressive, PERF = perfect.

The class of gapping clauses is equivalent to Sato's 'case element type' (Sato 1989), while the non-gapping clause class is comprised of his remaining relative clause types (see Sato (1989:9-12)). Whereas there exist a number of semantic subcategories for non-gapping clauses, the emphasis for this paper is placed on categorisation of the gap type of gapping clauses. That is, while non-gapping clauses are left as a generic set, gapping clauses are further divided into subcategories based on the case slot of the gap. This is based on the observation that for a random sample of relative clauses, an average of around 70% will be gapping, and only 30% non-gapping.[4]

The characterisation of relative clauses based on the presence of head gapping is well defined and unambiguous for a given interpretation of any relative clause. Thus, whereas semantically ambiguous relative clauses exist, each interpretation will be uniquely defined as either head-gapping or non-gapping. This is exemplified in sentence (3) below, in which the first gloss represents the interpretation of *keikaku* ('plan') as being gapped from the nominative case slot, while the second gloss is the non-gapping interpretation in which there is some contextually resolvable entity associated with the zero subject.

(3)	$\phi_{i/j}$	歴史	に	残る	計画$_i$
		rekisi	*ni*	*nokoru*	*keikaku*
	SUBJ	history	LOC	to remain-PRES	plan

'a plan which will go down in history' (*gap.*)

OR

'a plan for SUBJ$_j$ to go down in history' (*non-gap.*)

Note also that the definition of head gapping is based on full syntactic ellipsis of the gapped case slot, and that instances such as (4) below are defined as non-gapping, despite the head being semantically gapped as a subpart of the subject.

(4)	10月	から	新学期	が	始まった	北京大学
	10-gatu	*kara*	*singakki*	*ga*	*hazimatta*	*pekin-daigaku*
	October	from	new term	NOM	begin-PAST	Beijing University

'Beijing University, which began its new term in October'

Although the nominative case is instantiated by '*singakki*', the clause head can be seen to have been moved from a semantic part-gap in the nominative case, evidenced by:

(4a)	10月	から	北京大学	の	新学期	が	始まった
	10-gatu	*kara*	*pekin-daigaku*	*no*	*singakki*	*ga*	*hazimatta*
	October	from	Beijing University	GEN	new term	NOM	begin-PAST

'the Beijing University new term began in October'

Here, '***pekin-daigaku***' and '***singakki***' have been connected with the genitive marker in the nominative case slot. Clauses of this type represent a proper subset of Sato's 'indirect restrictive type' clauses.

3 A gap detection algorithm

Given the above definition for gapping and non-gapping clauses, we now present an algorithm to automatically categorise a given relative clause as either of these two types. The algorithm is based primarily on the inflectional features of the stem verb, and the semantic properties of the clause head. It returns the case slot of any gapped clause head, and terminates with failure for any non-gapped clause. The general gap detection algorithm is given in figure 2.

The algorithm takes as its input a VP NP construction in which the stem verb and complements of the clause body have been identified, including the case slot marker for each complement. In order to limit the reliance on semantic information, the accusative case slot is assumed to be uniquely defined by the *wo* marker ; this also acts to simplify the extraction process.[5] For the nominative case, on the other hand,

[4] Figures based on the random data set used for evaluation purposes.

[5] 'Travelling' usages of *wo* do not present a problem because they only occur with intransitive verbs, for which the algorithm does not access information regarding the accusative case slot.

```
$R_0$ Filter out any constructions with semantically illegal clause heads or stem verbs
   (see section 1 above).

$R_1$ IF the noun head is a non-gapping expression THEN the clause is non-gapping.
   RETURN *FAIL*

$R_2$ ELSE IF the noun head and stem verb are in a time relative construction THEN the
   clause is non-gapping.  RETURN *FAIL*

$R_3$ ELSE IF the noun head is a temporal expression AND the temporal case slot is
   uninstantiated THEN the head is gapped from the TEMPORAL case slot.  RETURN *TEMP*

$R_4$ ELSE IF the stem verb is transitive and not in the passive voice AND the ACC case
   slot is uninstantiated THEN the head is gapped from the ACCUSATIVE case slot.
   RETURN *ACC*

$R_5$ ELSE IF the NOM case slot is uninstantiated THEN the head is gapped from the
   NOMINATIVE case slot.  RETURN *NOM*

$R_6$ ELSE the relative clause is non-gapping.  RETURN *FAIL*
```

Figure 2: The gap resolution algorithm

any of the *ga*, *ha*, *mo* and *no* particles can collocate with the subject of a relative clause (see Tsujimura (1996:264-6) for details of the conditions required for this usage of the *no* particle). This brings about inherent problems because of the potential to use the *mo* and *ha* particles with the accusative case slot, but based on corpus analysis of the distribution of such usages, this effect produces only negligible noise.

For the passive voice, the system assumes that the stem verb is transitive, and that the object has been topicalised to the nominative case slot in an agentless passive construction.

Because of the reliance of the system on the transitivity of the stem verb, a dictionary is used to look up whether a given input stem verb is transitive or intransitive. Due to the limited coverage of this dictionary, a certain user-dependance is required for random input sets, in that the system asks the user to stipulate the transitivity of any stem verb not found in the dictionary.

3.1 Non-gapping expressions

Non-gapping expressions are defined as noun clause heads which are generally associated with a non-gapping, attributional interpretation of the relative clause construction. That is, the clause body represents an attribute of the clause head. One example of a non-gapping expression is *keikaku* ('plan') in sentence (3) above. Sentence (3) evidences the possibility for non-gapping expressions to be involved in gapping interpretations, but that this will not annul the non-gapping interpretation. Thus, in filtering all non-gapping expressions out as uniquely non-gapping expressions, the non-gapping interpretation will always be given preference for non-gapping expression clause heads.

Examples of non-gapping expressions are:

> *mokuteki* (目的 'purpose'), *ugoki* (動き 'movement/trend'), *hōsin* (方針 'direction/trend'), *kanzi* (感じ 'feeling'), *zizitu* (事実 'fact/truth')

Note that non-gapping expressions must constitute the full clause head to be able to guarantee non-gapping, and that the algorithm is based on a full match with an element in the dictionary of non-gapping expressions.

3.2 Time relative constructions

Time relative constructions are produced either by the clause head being a **time relative expression**, or by the head being a **time relative complex** and agreeing with the stem verb tense and aspect.

Examples of time relative expressions are:

sono-hi (その日 'that day'), *yokuzitu* (翌日 'the following day'), *tōzitu* (当日 'the day')

These can collocate with any stem verb inflection to generate a time relative interpretation.

Time relative complexes are generally produced by attaching a postfix to a phrase describing a time span. Instances of time relative complexes are:

1-kagetu-go (1カ月後 'one month later'), *nan-nitika-mae* (何日か前 'a few days before'), *2-okunen-mae* (2億年前 '200 million years before')

The two affixes currently treated by the system are *-go* ('after') and *-mae* ('before'). For *-go*, the stem verb must be in the simple past tense to produce a time relative construction, whereas *-mae* requires the simple present tense. If the head is a time relative complex but tense and aspectual requirements are not met, the clause is classified as a non-relative temporal construction, and the head is assumed to have been gapped from the TEMP case slot.

The effect of the tense and aspect of the stem verb in variously producing a time relative construction and a non-relative temporal construction, is illustrated by:

(5) 恐竜 が 住んでいた 約2億年前
kyōryū NOM *sunde-ita* *yaku-2-okunen-mae*
dinosaurs to live-PROG-PAST about 200 million years ago
'about 200 million years ago, when dinosaurs lived'

(6) 恐竜 が 住む 約2億年前
kyōryū NOM *sumu* *yaku-2-okunen-mae*
dinosaurs to live-PRES about 200 million years ago
'about 200 million years before dinosaurs lived'

Sentence (5) constitutes an absolute temporal construction, whereas the simple present tense in (6) leads to the production of a time relative construction.

The justification for the characterisation of time relative constructions as being non-gapping lies in the semantic incompatibility that exists between the time relative interpretation of the construction (*extra-clausal*), and the interpretation produced with the head in the TEMP case slot (*intra-clausal*). This can be seen in the glosses of sentences (5) and (6) above, in which (5) represents the clause head having been gapped from the TEMP case slot.

3.3 Temporal expressions

Temporal expressions consist of time-related NP heads which do not fit the definition of 'time relative expressions', and are assumed to have been moved from the TEMP case slot. They consist of **absolute temporal expressions, generic temporal expressions** and **non-relative temporal constructions**.

Absolute temporal expressions are of the type:

16-niti (16日 'the 16th'), *sakunen* (昨年 'last year'), *mainiti* (毎日 'everyday')

That is, they constitute the set of temporal expressions which are well defined within the context of the surrounding text.

Generic temporal expressions are of the type:

zidai (時代 'era'), *zikan* (時間 'time'), *nendo* (年度 'year/fiscal year'), *hi* (日 'day')

These express generic temporal categories and are semantically restricted by the clause body. They can be likened to lambda expressions in that they are ground TEMP case slot 'casts', without having the semantic extra-clausal and intra-clausal semantic incompatibility described for time relative constructions.

Non-relative temporal constructions are temporal constructions which involve a time relative complex head, but which do not fulfil the stem verb inflectional requirements of a time relative construction. Refer to section 3.2 above for details.

Note that there is a certain degree of reliance on the surrounding context as to whether a temporal expression is absolute or generic, in that most absolute expressions can be forced to take a generic reading.

However, within the confines of the current research, both the absolute and generic interpretations are assumed to have been gapped from the TEMP case slot, so the output of the algorithm is unaffected by this semantic phenomenon.

4 Evaluation

Evaluation of the algorithm was based on the EDR electronic corpus (EDR 1995). Relative clause constructions were extracted from the corpus, and the stem verb and its clause body complements determined through the use of a primitive grammar.

In order to compare the accuracy of the algorithm on both transitive and intransitive stem verbs, ergative pairs with case frames closely coinciding with those assumed in the algorithm were used, that is, the case frames given by:

$$\begin{array}{ll} \text{Subject} & \{\textit{ga,ha,mo,no}\} \\ & \text{NOM} \end{array} \left[\begin{array}{ll} \text{Object} & \textit{wo} \\ & \text{ACC} \end{array}\right] \begin{array}{ll} \text{Time} & \textit{ni} \\ & \text{TEMP} \end{array} \quad \text{Verb}$$

In the case of the intransitive ergate, the accusative case slot is deleted from the case frame.

The ergative pairs in question were *hazimaru* (始まる 'to begin' (*intrans.*)) /*hazimeru* (始める 'to begin' (*trans.*)), and *otiru* (落ちる/おちる 'to fall')/ *otosu* (落す/落とす/おとす 'to drop').

Subsequently, a set of 129 sentences containing relative clauses was randomly extracted from the corpus, for use as a baseline set.

Table 1 contains the results for each test set, partitioned across each case slot type, with the 'Others' entry referring to any case slot not modelled by the algorithm. The 'Overall accuracy' entry gives the overall proportion of heads correctly identified with their respective gap type, with the total number given in brackets. The '#' entry for each case slot gives the actual number of constructions of that type. For ACC case slot values given as 'N/A', the given verb is intransitive, and hence the ACC case slot will both never be instantiated and never be outputted by the algorithm.

The precision ($\mathcal{P}$) and recall ($\mathcal{R}$) are given for each case slot, with precision and recall defined respectively as:

$$\mathcal{P} = \frac{\textit{\# of instances of that type correctly identified by the algorithm}}{\textit{Total \# of instances categorised as that type by the algorithm}}$$

$$\mathcal{R} = \frac{\textit{\# of instances of that type correctly identified by the algorithm}}{\textit{Total \# of instances of that type}}$$

For figures marked as '—' in the table, a zero numerator has made the calculation of the given ratio impossible.

As can be seen from the results, the system performs marginally better on verbs which have a case frame closely coinciding with the basis of the algorithm, and returns its best performance for case frame coinciding intransitive verbs. The slightly stronger performance for the respective ergates of the *hazimaru*/*hazimeru* pairing over *otiru*/*otosu* was due to the second pair having an additional case frame associated with it, involving an extra TARGET case slot (e.g. 'to fall/drop down a hole'). This is reflected in the 'Others' row.

The worst performances were seen in both the precision and recall of the 'No gap' category for the random data set. However, given that this represents the default category, and that any case slots not modelled by the algorithm will produce noise, this is perhaps not surprising. The especially low recall is the result of the system having no access to the general context surrounding the particular relative clause, and thus not being able to give preference to extra-clausal antecedents to fill any ellipted case slots.

4.1 Limitations of the system

Clearly the algorithm is based on only the nominative, accusative and temporal case slots, and will fail for a clause head gapped from a case slot not corresponding to any of these slots. This was evidenced for the *otiru*/*otosu* ergative pair. However, as the evaluation of the algorithm accuracy on different input sets shows, such instances of relative clauses are uncommon and specialised to the particular verb type.

VERB:		*hazimaru*	*hazimeru*	*otiru*	*otosu*	random
Total sentences		184	91	64	38	129
Total clauses		185	91	64	38	147
Overall accuracy (*No. correct*)		99.5% (184)	94.5% (86)	95.3% (61)	94.7% (36)	90.5% (133)
Non-gapping exp. (R_1)	#	6	13	14	5	28
	$\mathcal{P}$	100%	100%	100%	100%	100%
	$\mathcal{R}$	100%	100%	100%	100%	100%
Time relative (R_2)	#	5	2	0	0	1
	$\mathcal{P}$	100%	100%	—	—	—
	$\mathcal{R}$	100%	100%	—	—	0%
TEMP (R_3)	#	27	9	2	0	4
	$\mathcal{P}$	96%	100%	80%	—	80%
	$\mathcal{R}$	100%	100%	100%	—	100%
NOM (R_4)	#	136	26	35	17	76
	$\mathcal{P}$	100%	96%	97%	89%	90%
	$\mathcal{R}$	99%	85%	100%	100%	100%
ACC (R_5)	#	N/A	38	N/A	11	17
	$\mathcal{P}$	N/A	90%	N/A	100%	94%
	$\mathcal{R}$	N/A	100%	N/A	100%	100%
No gap (R_6)	#	11	3	10	4	15
	$\mathcal{P}$	100%	100%	83%	100%	67%
	$\mathcal{R}$	100%	67%	100%	75%	53%
Others	#	0	0	3 (4.7%)	1 (2.6%)	6 (4.1%)

Table 1: The algorithm accuracy on the given data sets

Thus, any customisation of the algorithm to peripheral case slots would improve coverage only slightly, while potentially adversely affecting the precision of the system.

One further inherent problem associated with the algorithm is that a given verb may be associated with multiple case frames. Thus, while we suggest that the algorithm is fundamentally universal for all verbs and verb types, there will inevitably be slight variations stemming from case frame differences. This will also apply to the treatment of the passive voice with a particular stem verb. Whereas the use of the passive voice can represent passivisation, potential, respect, or rather a verb-specific idiom (see Tsujimura (1996:232-47) or Shibatani (1976) for details of the different usages), the current algorithm only considers the passive voice in the direct passive use and with the gap associated with the nominative case. As a result, the algorithm performs slightly less well with 'perception' verbs (e.g. *miru* (見る 'to see') and *kangaeru* (考える 'to think')), which display idiosyncratic patterns of usage of the passive voice.

Similarly, 'empathy' (Kuno and Kaburaki 1977) receives no treatment currently, justified by its sparseness of use in text-based corpora. This presents a coverage problem, especially for any application of the algorithm on speech data.

One difficulty with basing the algorithm on purely lexical and syntactic filtering is that there is no way to deal with lexically ambiguous clause heads. Examples of such heads are the kanji '方' and '通り'. While '通り' has a unique reading as *tōri*, it can either be used as a clause-level grammatical construction, or as a common noun meaning 'road'. As such, the first usage should be filtered out as a non-relative clause construction, whereas the second noun interpretation will produce a relative clause construction and should be retained. Given the existing simplified grammar, there is no possibility of detecting grammatical constructions at the clause level, and these must instead be manually removed from data. Similarly for '方', the *hō* reading can either be used as part of a fixed expression or as a pronoun ('one'), producing the same ambiguity as for *tōri*. There is also an additional reading for '方' in *kata*, meaning 'person' (*honorific*).

4.2 Future work

Future work will consist of making the handling of the passive voice more sophisticated and including resolution of empathy, as described in section 4.1 above. Additionally, work is required to filter out idiomatic relative clause uses, as these generally represent non-gapping relative clauses with deleted case slots. This process can be faciliated by making use of the well defined relationship that generally exists between the stem verb and clause head.

Additionally, the system should be extended to include the use of semantic filters to estimate the compatibility of the clause head with a particular case slot. Clearly, weighting should be used to integrate the syntactic constraints proposed here with semantic constraint information.

The primitive grammar used to extract relative clause constructions from the EDR corpus also requires reworking, in order to differentiate between clause-level usages and genuine relative clauses, as was seen for the case of *tōri* above.

5 Conclusion

We have proposed a simple algorithm for gap resolution of Japanese relative clause constructions. The algorithm is realised by a reliance on syntactic and lexical information, but based partially on semantic analysis. It determines whether the head is gapped within the clause body, and for any gapped instances, returns the case slot of the gap.

The algorithm was evaluated for two well-defined ergative pairs, for which it returned a median accuracy of around 95%. On a random sample set, the algorithm achieved an accuracy of around 90%. Case slots specific to the given stem verb and not modelled by the algorithm, proved to be the main source of errors.

Future work will consist of improving handling of the passive voice, expansion of the algorithm to include empathy, and addition of semantic restrictions. The applications of the given algorithm for general discourse processing is also an area requiring further exploration.

References

EDR, 1995. *EDR Electronic Dictionary Technical Guide*. Japan Electronic Dictionary Research Institute, Ltd.

HALLIDAY, M. A. K. 1994. *An introduction to Functional Grammar*. Edward Arnold, 2nd edition.

KITAGAWA, C. 1982. Topic constructions in Japanese. *Lingua* 57.175–214.

KUNO, S., and E. KABURAKI. 1977. Empathy and syntax. *Linguistic Inquiry* 8.627–72.

SATO, R., 1989. *Research relating to the semantic analysis of Japanese attributive clauses.* Master's thesis, Tokyo Institute of Technology.

SHIBATANI, M. 1976. *The Grammar of Caustative Constructions*. Academic Press: New York.

TSUJIMURA, N. 1996. *An introduction to Japanese linguistics*. Blackwell.