Semantic Verb Classes in the Analysis of Head Gapping in Japanese Relative Clauses

Timothy Baldwin, Takenobu Tokunaga and **Hozumi Tanaka**
Tokyo Institute of Technology
{tim,take,tanaka}@cs.titech.ac.jp

Abstract

This paper describes an attempt to identify case gapping instances of Japanese relative clauses, and disambiguate the case slot from which the gapping occurred. The method utilised relies principally on surface syntactic pattern matching, combined with adjunct-based verb classification and minimal semantic analysis of the head. Experimentation produced an 90% accuracy in identifying and disambiguating case gapping instances, and 95% accuracy for the isolated task of gapping type disambiguation.

1 Introduction

Relative clauses in Japanese display uniform syntactic structure in performing a remarkable range of distinct semantic roles. Past research has focused on describing and classifying this diversity, but no real effort has been made to automate the classification process or resolve the semantic relationship between the relative clause body and head. While this paper does not claim to have filled this gap, it does propose one means of going about the classification process, distinguishing between gapping and non-gapping clauses, and resolving the head gap for gapping clauses.

2 Definitions

2.1 Relative clauses in Japanese

In Japanese, relative clauses are syntactically identical to tensed matrix clauses[1], but are immediately proceeded by the modified noun phrase (NP). This basic VP NP structural requirement is qualified to exclude relative clauses headed by "formal nouns" (see (Shibatani, 1978, pp 69-70), (Kuno, 1973, pp 137-142)), and instances where the scope of the relative clause body modification extends to the superordinate clause level.

Throughout this paper, we will refer to the modified NP head as the 'head', the modifying clause (VP) as the 'clause body', and the combined VP NP construction as a 'relative clause construction'.

2.2 Gapping relative clause types

One attempt to describe the full spectrum of relative clause types was made by Sato, who devised a hierarchy of 5 main classifications, including the '*case element*' and '*indirect restrictive*' gapping types (Sato, 1989, pp 8-12). Our research principally focuses on these two types, which account for a combined proportion of over 85% of a typical selection of relative clauses[2], and are the only types to involve case gapping.

Case element clauses are defined as those in which the head has been 'gapped' from within the case frame subtended by the main verb of the clause body. In the case of Japanese, the gapping process is characterised by a full case slot being moved from within the clause body, involving the deletion of the case marker associated with the gapped case element and effectively removing any surface indication of the gap identity. Let us consider this process through the following example of a case element clause[3]:

(1)	*kyou*	ϕ_i	ϕ_j	*katta*	*hon*$_j$
	today	SBJ	DO	to buy-PAST	book
	'the book which (I) bought today'				

The syntactic structure of the original Japanese relative clause construction closely parallels that of the English gloss, in that the head of *hon* has been gapped from the direct object position. However, unlike the English gloss, the Japanese relative clause contains both the direct object position gap and an ellipted subject, with no surface marking of the direct object case slot as the gap for the head. As such, (1) is syntactically ambiguous between the subject

[1] As such, verb, adjectival verb and adjectival phrases can all form relative clauses (see (Kanzaki, 1997) for details of adjectival usages), but for the purpose of this research, relative clauses are assumed to be restricted to verb phrases (VPs).

[2] The figure of 85% is based on a sample of 3784 distinct relative clause construction instances, of which 3306 were case element or indirect restrictive relative clauses, 71 were idiomatic usages, and the remaining 407 were non-gapping.

[3] The following case marker nomenclature is used in glosses: NOM = nominative, ACC = accusative. Deep case markers are indicated by: SBJ = subject, DO = direct object, IO = indirect object, TOP = topic, LOC = locative, TEMP = temporal, and EXP = experiencer. "ϕ" is used to indicate zero anaphoric verb complements.

In Proceedings of the Natural Language Processing Pacific Rim Symposium '97
(NLPRS'97), Phuket, Thailand, pp. 409-14.

and direct object case slots as to the identity of the head gap.

In the case of indirect restrictive relative clauses, the head can be seen to constitute the 'anchor' for a case filler realised in the clause body. To illustrate this with an example:

(2) *pēzi-ga* *otiteiru* *hon*
page-NOM to fall-PRES-PROG book
'a book which has missing pages'

Based on the English gloss, it would appear that the syntactic mechanism at play here is identical to that for (1), and that *hon* is gapped from the subject case slot. However, as is evident from the structural analysis given for (2), the subject case slot is instantiated by *pēzi*, which is then anchored by *hon*. As pointed out by Sato, the nominative case filler can be considered to be restricted by the head, such that *pēzi* represents a partially gapped version of the restrictive genitive NP *hon-no-pēzi* ('pages of a book'). Alternatively, the head can be considered as the 'major subject' for the relative clause (Tateishi, 1994, pp 28-38), and hence as having been gapped from the *topic* case slot.

2.3 Non-gapping relative clauses

The difficulty of the case gapping disambiguation process lies in the fact that multiple zero anaphoric case slots can coexist, as is the case for (1), and that the existence of ellipted case slots does not necessarily produce a case element clause. Indeed, ellipted case slots can refer to any of interclausal, intraclausal or deictic elements.

To illustrate this point, let us consider an example of a non-gapping relative clause:

(3) *ϕ* *hon-wo* *katta* *riyū*
SBJ book-ACC to buy-PAST reason
'the reason (I) bought the book'

Here, the syntactic content of the relative clause is similar to that for (1), but the head has been replaced with the functional noun *riyū*, producing a non-gapping relative clause. That is, the ellipted subject case slot refers to a context-resolvable actor, and is not a gap for the head. Note, however, that a functional noun head is neither sufficient nor necessary to produce a non-gapping relative clause, and that the existence of ellipted case slots leads to potential ambiguity in classifying the clause in terms of the gapping/non-gapping relative clause paradigm.

3 Case frame representation and verb classification

The problem targetted in our research is the identification of case element and indirect restrictive relative clauses, and the case frame-based analysis thereof. For this purpose, we require a case frame representation for the main verb, which is then combined with inflectional and verb class data.

3.1 Case valence dictionary

The case frame and verb class content of the case valence dictionary used in our research is based largely on the valency dictionary developed by NTT for Japanese-to-English machine translation (Ikehara et al., 1997).

To avoid consideration of verb sense disambiguation, a unique default entry is given for each root verb, containing a generic complement-based case frame. Low level preference heuristics are introduced into the case frame by reverse ordering the component case slots in terms of their default propensity to gapping.

In order to complement the default entry for each root verb, high levels of fixed expressions and idioms have been introduced into the dictionary, largely from the NTT valency dictionary. Fixed expressions differ from default entries in that they have a set of case slots which must match in case filler content with the corresponding case slot in the system input. They include an analysis of those instantiated case slots which can be gapped as the head, information which was not available directly from the original representations.

'Conditional instantiated' entries are also contained for certain verbs, to capture well-defined verb senses which behave differently to the default sense, but in order to be triggered, require the instantiation of an arbitrary set of case slots disjunctive in content with the default case frame. That is, conditional instantiated entries are roughly equivalent to fixed expression entries, except that simple case slot instantiation, rather than case filler correspondence, is required to produce that meaning.

Verb class information is given for each dictionary entry to implicitly model peripheral adjuncts. That is, the full case frame content of the root verb is made up of the complements provided in the dictionary case frame, combined with the implicit representation of adjuncts given by verb class information.

3.2 Verb-based lexical ambiguity

In order for the system to guarantee at most one output for each verb root, fixed expressions are given preference over conditional instantiated entries, which in turn take precedence over the default entry. Additionally, fixed expression entries have been pre-editted to ensure that no lexical overlap exists between the fixed case element content of entries for a given verb root.

Despite this guarantee of at most one output for each verb root, inflectional and lexical ambiguities in Japanese can lead to multiple verb root correspondence for a given verb input. This leads to the potential for multiple dictionary entries being triggered, the scope of which is only restricted by the entry type preferences described above. In such cases, the system processes all successfully parsed dictionary

entries in parallel and the resultant set of candidate solutions is returned as the system output.

4 The algorithm

The basic algorithm used for gapping resolution consists of three declarative rule sets, the *inflectional-based*, *verb class* and *default* rule sets. These are applied linearly until a component resolution rule is triggered.

4.1 The inflectional-based rule set

The inflectional-based rule set is characterised by verb and head-based rules being interleaved with highly specific inflectional heuristics. Due to space limitations, the reader is asked to refer to (Baldwin et al., 1997b, p 279) for the basic details of the inflectional heuristics.

First, the system treats idioms (1a) and relative clauses headed by "non-gapping expressions" (1b). Clauses headed by non-gapping expressions are such that the clause can be predicted with high confidence to be non-gapping (Baldwin et al., 1997a, p 4)[4]. If a full match for the head is not obtained, the rule set proceeds to determine if the main verb inflection is of a type which generally brings about a transformation in the basic structure of the verb-based frame (principally passive and causative inflection). That is not to say that all case slots of these inflectional types are handled at this stage, but rather those instances of heads gauged to have been gapped from transformed case slots are filtered off, with untransformed case slot instances left for the verb class rule set.

A second function of the first rule set is to extract a series of well-defined relative clause usages which closely inter-relate with the inflectional heuristics contained in the rule set. The incorporation of these clause types at this premature stage of processing is related to the optimum ordering for the inflectional heuristics with respect to the handling of the various verb classes.

Specifically, the two verb class sets which are incorporated here are the *excluding* and *copulative/conjoining* class sets. *Excluding* (1c) consists uniquely of the verb *nozoku*, for simple tense usages with only the accusative case slot instantiated.

(4) *nitiyou-wo* *nozoku* *mainiti*
Sunday-ACC to exclude-PRES everyday
'everyday, excluding Sundays'

While the head can be thought of as having been gapped from the 'from' case slot, this case slot is not syntactically realiseable for the above 'excluding' sense of *nozoku*. As such, excluding clause instances are treated as being non-gapping.

The *copulative/conjoining* verb classes (1d), on the other hand, are much more orthodox in their behaviour, being incompatible with all locative case slot types, and not readily associating with temporal usages. Examples of conjoining verbs are *uwamawaru* ('to exceed') and *kanren-suru* ('to relate to')

Discussion of the treatment of the temporal case slot (1e) is given in (Baldwin et al., 1997a, pp 4-6).

4.2 The verb class rule set

The verb class rule set is partitioned according to verb class, and interfaces with the case frame for the main verb in the relative clause. Basically each verb class-customised rule subset is inbuilt with adjunct-based information, and uses the semantic content of the head[5] to determine whether it should be mapped onto one of these peripheral case slots, or alternatively one of the case frame slots. The version of the verb case frame the system receives has been modified by excluding those verb slots which are instantiated within the clause body, and the system uses the ordering of the resultant case frame, combined with the semantic content of the head, to heuristically choose the most likely (uninstantiated) case slot; this function is performed by `CASE_RESOLVE` in the algorithm. It is important to bear in mind that verb entries are generally attributed with multiple verb classes, and that a single entry can 'fall through' different applicable rule subsets before finally triggering a rule.

Verb classes included in the verb class rule set can be described as being of three main types: *existential*, *relational* and *movement*.

Existential verbs (2a) are commonly associated with compatibility with the locative case slot, realised through the *ni* marker. 'Stative' verbs form a proper subset of existential verbs, and are additionally compatible with the experiential dative case slot. Examples of existential and stative verbs are, respectively, *sumu* ('to live/inhabit') and *aru* ('to have/be').

Relational verbs are characterised by relating a source and target entity. In the case of inter-personal relational verbs (2b), both entities are generally human, whereas generic relational verbs (2c) are associated with a broader range of arguments. The principle mechanism at work with both types is that, except under exceptional circumstances, the source entity must be instantiated to trigger the target entity. That is, the target entity is included as a verb constituent in the case frame, but indicated as requiring activation either from the similarly marked source entity, or indirectly from the semantics of the head.

'Empathy' verbs form a proper subset of inter-personal relational verbs, and are such that the emphatic focus on the source object is sufficiently high that the target entity can be activated from a syntactically unrealised source entity (see (Iida, 1996,

[4] Non-gapping expressions comprise a subset of functional nouns (see section 2.3)

[5] Analysed through the use of the NTT semantic dictionary (Ikehara et al., 1993), (Ikehara et al., 1997)

THE INFLECTIONAL-BASED RULE SET

1a: IF (the construction is **idiomatic**) RETURN *IDIOM*;

1b: IF (the head is a **non-gapping expression**) RETURN *NE*;

1c: IF (excluding clause instance) RETURN *FROM_EX*;

1d: IF (copulative or conjoining main verb)

IF (**subject** case slot uninstantiated) RETURN *SBJ*;

ELSE GOTO DEFAULT;

1e: IF (head constitutes **temporal** case slot) RETURN *TEMP*;

THE VERB CLASS RULE SET

2a: IF (existential main verb)

IF (**locative** head AND uninstantiated **locative** case slot) RETURN *LOC*;

ELSE IF (uninstantiated **subject** case slot) RETURN *SBJ*;

ELSE IF (stative main verb) RETURN *EXP*;

2b: IF (inter-personal relational main verb)

IF (**autonomous** head)

IF (empathy main verb AND relative clause is tensed or has locative case element) RETURN *IO*;

ELSE IF (uninstantiated **target** case slot α) RETURN α;

ELSE RETURN *IO*;

ELSE IF (uninstantiated **locative** case slot) RETURN *LOC*;

ELSE IF (uninstantiated **source** case slot β) RETURN β;

ELSE IF (CASE_RESOLVE identifies an uninstantiated case slot γ) RETURN γ;

2c: IF (generic relational main verb)

IF (instantiated **target** case slot AND uninstantiated **source** case slot α) RETURN α;

ELSE IF (CASE_RESOLVE identifies an uninstantiated case slot β) RETURN β;

2d: IF (distal movement main verb)

IF (**autonomous** head AND uninstantiated **subject** case slot) RETURN *SBJ*;

ELSE IF (**locative** head AND uninstantiated **destination locative** case slot) RETURN *TO*;

2e: IF (travelling main verb)

IF (**autonomous** head AND uninstantiated **subject** case slot) RETURN *SBJ*;

ELSE IF (**locative** head AND uninstantiated **'through' locative** case slot) RETURN *THROUGH*;

THE DEFAULT RULE SET

3a: IF (**autonomous** head) THEN

IF (uninstantiated **subject** case slot) RETURN *SBJ*;

ELSE IF (**first person pronoun** head) RETURN *TOP*;

3b: IF (CASE_RESOLVE identifies an uninstantiated case slot α) RETURN α;

3c: IF (**locative** head AND generic action main verb) RETURN *LOC*;

3d: ELSE DEFAULT:

IF **non-abstract** head RETURN *TOP*;

ELSE RETURN *NG*;

Figure 1: The combined rule sets

pp 326-44) for a fuller account of the intricacies of emphatic focus). Unfortunately, it is often difficult to unambiguously designate which of the source and target case slots the head has been gapped from, except in cases where the clause is temporally or locatively grounded. This is one drawback of the current algorithm, in that it cannot weight multiple solutions according to confidence, but simply returns a unique solution.

Heads which activate the target case slot are those which inherently refer to a target entity, the most common of which is *aite* ('opponent/partner').

Distal movement verbs (2d) are similar in nature to existential verbs, except that the *ni* locative marker is commonly interchangeable with *e* in the case frame. In the case of travelling verbs (2e), the locative case slot marker is not *ni* as for existential and distal movement verbs, but *wo* in the travelling sense. *iku* ('to go') is one example of a distal movement verb, and *tōru* ('to travel/pass through') is a travelling verb.

4.3 The default rule set

In the event that the system does not return a solution for either of the first two rule sets, it applies the third rule set as a default, which is guaranteed to produce an output. Firstly, the system attempts to map autonomous heads onto the subject case slot, and failing this, first person pronouns are assumed to correspond to the clause topic (3a). Note that the clause topic analysis for first person pronouns corresponds to the identification of indirect restrictive relative clauses.

Failing the identification of a complement case slot gap for the head (3b), the system attempts to map the head onto the *de* action locative case slot (3c), associated with 'generic action' verbs.

For the system to reach the final 'DEFAULT' rule subset (3d), all complement case slots must be instantiated, discluding the relative clause from being of the case element type. Thus, this rule set is primarily used to differentiate between indirect restrictive usages ('TOP') and non-gapping relative clauses ('NG').

5 Evaluation

Evaluation of the system was carried out based on a test set of relative clause constructions extracted from the Japanese EDR corpus (EDR, 1995). The test set was classified according to both the verb class contents and inflection of the main verb, to verify the accuracy of the above algorithm on each verb class type, and to test the accuracy of the inflectional heuristics. Additionally, fixed expressions were identified to compare the overall system performance based on default case frames, and that for fixed expressions. For reasons explained in section 3.2, the system can produce multiple outputs for a single input, and for these cases the system output was adjudged as being correct if the correct analysis was contained in the set of candidate solutions.

The system accuracy was analysed according to the overall accuracy for each test set, and also the accuracy on only gapping examples[6]. For each class of rule set outputs given in 1, the actual number of instances of that class for the different data sets is given as '#', with data then analysed in terms of precision (P) and recall (R) (based on the standard definitions). Cases where a zero denominator has made either of these values incalculable are indicated in the results as 'N/A'.

Table 1 first details the overall performance of the system on fixed expressions and effectiveness of the inflectional heuristic component of the inflectional-based rule set, including analysis 'with' and 'without' these heuristics. Data for excluding and copulative/conjoining verbs is combined in the 'inflectional verb types' column, completing evaluation of the inflectional-based rule set.

Given that the rules making up the inflectional-based rule set are those for well-defined, self-contained verb types, and relatively easily predictable syntactic phenomena, it is unsurprising that all these results are comparatively high. The results 'with' and 'without' the inflectional heuristics are gratifying, and suggest that we have more than halved the error margin for affected clauses and significantly improved the precision for indirect restrictive clauses ('TOP').

Results for the three main verb types treated in the verb class rule set (existential, relational and movement verbs, respectively) are given next, along with an indication of the overall system performance. On a dry run, the system seems to perform at around 90% accuracy, but when errors for non-gapping clause misidentification are removed from this figure, the performance on gapping relative clause constructions seems to be as high as 95%, reaching 96.7% for relational verbs. One clear weakness is the detection of locative instances, particularly for movement verbs (represented as the principal component of the 'Other case slots' row).

Comparing the overall results for the verb class-defined test sets with those for fixed expressions, the benefits of the finer sense granularity derived through fixed expressions become apparent, especially when considering the 98.0% accuracy for gapping instances. That is, through use of fixed expression verb entries, we are able to reduce the overall system error by more than half.

Perhaps the most disappointing result comes in the overall evaluation of indirect restrictive clauses (precision 60.4%), although the significantly higher figures for each of the verb classes are reassuring.

[6] In terms of the tags returned from the algorithm, all *NE* and *NG* instances were excluded from the data, as well as adjectival head-based constructions and time relative constructions (Baldwin et al., 1997a, pp 4-5).

TEST SET: (No. of clauses)		*Fixed expressions* (160)	*Inflectional heuristics* (751) w/o	with	*Inflectional verb types* (1099)	*Exist* (137)	*Relation* (513)	*Movement* (276)	Overall (4222)
Accuracy (excl. non-gapping)		**90.0%** (98.0%)	**62.1%** (68.8%)	**87%** (95.8%)	**97.8%** (99.3%)	**95.6%** (96.3%)	**91.8%** (96.7%)	**86.6%** (94.9%)	**89.5%** (95.0%)
TEMP	#	4	14		1	4	13	3	107
	P	100%	70.0%	100%	100%	100%	100%	100%	100%
	R	100%	100%	100%	100%	100%	100%	100%	96.3%
SBJ	#	78	501		944	86	308	151	2895
	P	87.6%	86.1%	90.0%	98.8%	98.8%	91.0%	86.4%	92.6%
	R	100%	64.1%	97.2%	99.8%	96.5%	99.0%	96.7%	97.9%
DO/EXP	#	38	57		0	9	24	22	280
	P	90.2%	20.2%	72.9%	*N/A*	90.0%	88.9%	84.0%	78.1%
	R	97.4%	87.7%	89.5%	*N/A*	100%	100%	95.5%	83.9%
IO	#	0	2		1	1	61	2	74
	P	*N/A*	*N/A*	100%	*N/A*	*N/A*	92.9%	100%	80.3%
	R	*N/A*	0%	50.0%	0%	0%	85.2%	50.0%	71.6%
Other gapping case slots (LOC/TO/..)	#	1	7		110	24	19	13	222
	P	100%	30.0%	75.0%	98.2%	92.3%	100%	71.4%	81.0%
	R	100%	85.7%	85.7%	100%	100%	100%	76.9%	84.2%
TOP (*indirect restrictive*)	#	14	25		14	6	18	14	80
	P	100%	44.4%	60.6%	75.0%	100%	100%	92.9%	60.4%
	R	92.9%	16.0%	80.0%	64.3%	83.3%	77.8%	92.9%	68.8%
NG/NE (*non-gapping*)	#	25	145		29	7	70	71	564
	P	100%	92.2%	89.0%	61.9%	75.0%	89.8%	90.2%	83.6%
	R	48.0%	49.0%	55.9%	44.8%	85.7%	62.9%	64.8%	55.9%

Table 1: The algorithm accuracy on different input types

6 Conclusion

Through use of case frame information and verb classes, we have produced a working system which analyses Japanese relative clause constructions with an overall accuracy of around 90%, with analysis of gapping clauses 95% accurate. The driving mechanism utilised to achieve this figure is syntactic, with minimal reliance made on a thesaurus in determining the semantic content of the head.

Admittedly one of the greatest drawbacks of the system in its current form is its relatively limited ability to recognise non-gapping clauses, and this presents a major area for future improvement. Additionally, more work is required to mechanically distinguish between the full range of relative clause types described by Sato.

Acknowledgements

The authors would like to thank the staff of NTT for making available their considerable electronic resources, and particularly Francis Bond for valuable comments on this paper.

References

T. Baldwin, H. Tanaka, and T. Tokunaga. 1997a. Analysis of head gapping in Japanese relative clauses. In *Information Processing Society of Japan SIG Notes*, volume 97, no. 4, pages 1–8.

T. Baldwin, T. Tokunaga, and H. Tanaka. 1997b. Syntactic and semantic constraints on head gapping in Japanese relative clauses. In *Proc. of the Third Annual Meeting of the Japanese Association for Natural Language Processing*, pages 277–80.

EDR, 1995. *EDR Electronic Dictionary Technical Guide*. Japan Electronic Dictionary Research Institute, Ltd. (In Japanese).

M. Iida. 1996. *Context and Binding in Japanese*. CSLI Publications.

S. Ikehara, M. Miyazaki, and A. Yokoo. 1993. Classification of language knowledge for meaning analysis in machine translation. *Transactions of the Information Processing Society of Japan*, 34:1692–1704. (In Japanese).

S. Ikehara, M. Miyazaki, A. Yokoo, S. Shirai, H. Nakaiwa, K. Ogura, Y. Ooyama, and Y. Hayashi. 1997. *Nihongo Goi Taikei – A Japanese Lexicon*. Iwanami Shoten. 5 volumes. (In Japanese).

K. Kanzaki. 1997. Lexical semantic relations between adnominal constituents and their head nouns. *Mathematical Linguistics*, 21(2):53–68. (In Japanese).

S. Kuno. 1973. *Nihon Bunpō Kenkyū*. Taishukan. (In Japanese).

R. Sato. 1989. *Nihongo no Rentai-shūshoku-setsu no Imi-kaiseki ni-kansuru Kenkyū*. Master's thesis, Tokyo Institute of Technology. (In Japanese).

M. Shibatani. 1978. *Nihongo no Bunseki*. Taishukan. (In Japanese).

K. Tateishi. 1994. *The Syntax of 'Subjects'*. CSLI.