Syntactic and Semantic Constraints on Head Gapping in Japanese Relative Clauses

Timothy Baldwin, Takenobu Tokunaga, Hozumi Tanaka
Tokyo Institute of Technology

{tim,take,tanaka}@cs.titech.ac.jp

0 Introduction

This paper represents the continutation of research into the identification of the relationship between the head of a relative clause and the clause body, for Japanese. Unlike many languages, a 'trace' of the gap associated with the head of a gapping relative clause is not lexically marked in Japanese, and gapping and non-gapping clauses are not explicitly distinguished. Our current research is focused on distinguishing between gapping and non-gapping relative clauses, and determining the identity of the gapped case slot for gapping clauses.

1 Definitions

For the purposes of this paper, a relative clause can be considered simply as a noun phrase (NP) modified prenominally by a clause (VP), where the modified NP is refered to as the 'clause head' or 'head', and the modifying clause as the 'clause body'. NP complexes which fulfill this simple (VP NP) structural constraint but are preculded from the definition of relative clauses, include instances where the scope of the clause body modification extends to the clause-level, and NP heads which cannot exist independently of a modifying clause body (see (Baldwin *et al.* 1997:2) for further details).

Head gapping is then defined as the process whereby the clause head can be considered to have been 'gapped' or 'moved' from an ellipted case slot within the clause body. Considering example (1) below, the 'user' clause head can be seen to have been gapped from the nominative case slot of the clause body stem verb, 'to be satisfied'. As such, this provides an example of head gapping, while example (2) provides an example of a non-gapping relative clause.

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

(2) ϕ 勝つ 意志
katu *isi*
SUBJ to win will
'the will to win'

It is important to note here that, whereas relative clauses in English implicitly incorporate head gapping, our definition of 'relative clause' for Japanese includes both NP complexes that involve head gapping, and those for which the clause body and clause head are syntactically independent. This difference in the definitions for the two languages is evidenced in the lack of an English gloss for example (2) which incorporates a relative clause.

2 The original gap resolution algorithm

The original gap resolution algorithm proposed in our previous paper (Baldwin *et al.* 1997) is indicated in figure (1) as those rules marked with $\mathbf{O}_m$, and is contrasted with the revised algorithm and rules $\mathbf{R}_n$. The respective sequentiality of the two rule numbering systems evidences the adherence of our revised algorithm to the basis of the original framework, with the addition of both semantic and syntactic-based rules to fine tune the classification process. Due to this reliance on the essence of the original algorithm, we describe that algorithm here.

2.1 A description of the original algorithm

The original algorithm relies primarily on a transitivity judgement for the stem verb of the relative clause, and analysis of the surface case markers contained within the clause body. The first stage of the determination process involves preprocessing a subset of non-gapping relative clauses, through the use of a static dictionary containing head instances generally associated with non-gapping clauses ($\mathbf{O}_1$). Time-related relative clauses are then filtered off, based on simple

head recognition/classification templates and the stem verb tense ($\mathbf{O}_{2,3}$). Of these, time relative constructions represent instances where the temporal context of the clause head is determined relative to that of the clause body ($\mathbf{O}_2$) ; temporal constructions, on the other hand, represent instances of the head being well-defined within the general context of the surrounding text (and hence semantically unresticted), or the head being a generic temporal expression ground by the restriction of the clause body ($\mathbf{O}_3$).[1]

The default mechanism used to classify remaining instances of relative clauses preferentially models the gapped case slot as the accusative case slot, or failing this, the nominative case slot ($\mathbf{O}_{4,5}$, corresponding to a generic version of $\mathbf{R}_{14}$). Failed case slot allocation variously results from: incompatibility in the voice of the stem verb, the instantiation of a given case slot, or the transitivity of the stem verb. Note also that the accusative case slot is assumed to conicide uniquely with the *wo* (を) case marker.

In cases where the default classification process in $\mathbf{O}_4$ and $\mathbf{O}_5$ fails, the relative clause is assumed to be non-gapping ($\mathbf{O}_6$).

2.2 Limitations

The original algorithm is limited in the following respects:

1. **Idioms**
   - Not treated
2. **Over-generalised case frame**
   - Reliance on one generic case frame for each of the sets of transitive and intransitive verbs, with transitive verbs assumed to be uniquely associated with the *wo* case marker for the accusative case slot.
3. **Passive voice**
   - The clause head is assumed to have been gapped from the nominative case slot for passive stem verbs, and in the case of an instantiated nominative case slot, the relative clause is classified as being non-gapping.
4. **Empathy effects resulting from the stem verb**
   - Not treated
5. **Causative verb inflection**
   - Not treated
6. **Locative gapping**
   - Not treated

[1] Refer to (Baldwin *et al.* 1997:4-6) for specific details and examples of both time relative constructions and temporal constructions.

3 Extensions to the original algorithm

The revised version of the algorithm is detailed in figure 1, and the various extensions to the original algorithm are described below.

3.1 Idioms

An idiom dictionary was produced, consisting of templates of relative clause-based idiomatic usages. For idiom recognition, the dictionary requires a full match of all case slots within the clause body, the stem verb (including inflection), and the semantic head of the clause head ($\mathbf{R}_1$). For purely verb-based idiomatic usages, the clause head is marked as being unlimited in scope, and the dictionary match is based simply on the content of the stem verb.

3.2 Customised case frame modelling

A syntactic case frame dictionary was developed, in which each verb is associated with a unique ordered list of case slot markers, with parallel case frames represented by multiplicity of case marker candidates for the corresponding case slots. Temporal and locative case slots are not included in case frame representations, in an attempt to minimise potential differences between multiple case frames for a given stem verb. Rather, all verbs are assumed to be potentially compatible with these two gapping types, and rules to model them are incorporated into the gap resolution algorithm ($\mathbf{R}_5$ and $\mathbf{R}_{15}$ respectively).

3.3 Passive stem verbs

The basis of the original algorithm with respect to the passive voice is retained, in that the default rule attempts to instantiate the clause head with the nominative case slot ($\mathbf{R}_3$). However, as an extension to the original algorithm, the resolution procedure is allowed to continue through the remaining rules in the event of an instantiated nominative case slot.

3.4 Empathy

A first person pronoun detection template is combined with the NTT semantic dictionary (Ikehara *et al.* 1993) to filter out human-referring clause heads. This information is then coupled with a dictionary of 'empathy-type' verbs in rules $\mathbf{R}_6$-$\mathbf{R}_9$, whereby preference is given to the nominative case slot for 'neutral' (non-empathy) stem verbs and first person pronomial clause heads, but to the accusative case slot otherwise. Verbs defined as being 'empathy-type' are transitive and potentially involve both a human subject and object, but with the

FOR EACH PARSEABLE STEM VERB ENTRY:

R_1 IF the construction is idiomatic RETURN *IDIOM*

O_1/R_2 IF the clause head is a non-gapping expression RETURN *NE*

R_3 IF passive stem verb and uninstantiated nominative case slot RETURN *NOM*

O_2/R_4 IF clause head and stem verb comprise a time relative construction RETURN *RELT*

O_3/R_5 IF noun head is a temporal expression AND temporal case slot is uninstantiated RETURN *TEMP*

R_6 IF clause head is human

R_7 IF nominative case slot is uninstantiated

R_8 IF 1st person clause head OR NOT 'empathy-type' stem verb RETURN *NOM*

R_9 ELSE IF 1st person clause head RETURN *NON_GAPPING*

R_{10} IF stem verb is causative

R_{11} IF nominative case slot is uninstantiated RETURN *NOM*

R_{12} IF *wo* case slot is uninstantiated RETURN *ACC*

R_{13} REPEAT Work through the list of case markers for stem verb UNTIL find unistantiated case slot CS_m RETURN CS_m

$O_{4,5}/R_{14}$ REPEAT Work through list of case markers for stem verb UNTIL find uninstantiated case slot CS_n RETURN CS_n

R_{15} IF clause head is locative RETURN *LOC*

O_6/R_{16} RETURN *NON_GAPPING*

Figure 1: The revised gap resolution algorithm

focal 'empathy' (Kuno and Kaburaki 1977) being on the subject. Examples of such verbs are *au* (会う 'to meet') and *wakareru* (別れる 'to leave/break up with').

3.5 Causative inflectional

Causative inflection is modelled similarly to the passive voice, in that the nominative case slot is assumed to be most strongly binding. Unlike the passive voice, however, an extra *causee* element is generated, which is assumed to coincide with the *wo* case slot ($\mathbf{R}_{10}$-$\mathbf{R}_{13}$).

3.6 Locative case slot

The NTT semantic dictionary (Ikehara *et al.* 1993) is used to filter out locative clause heads, which are then mapped onto the locative case slot ($\mathbf{R}_{15}$). Due to the relative infrequency of instances of the clause head being gapped from the locative case slot, this rule is maintained as a default.

4 Evaluation

Both the original and revised algorithms were applied to the analysis of relative clauses contained in four separate sentence sets extracted from the EDR corpus (EDR 1995). During the extraction phase, each relative clause was analysed with a primitive *bunsetsu*-style grammar to derive its clause body complements, stem verb, and the semantic head of the clause head. A gapping judgement was then manually determined for each relative clause.

Sentences in the first two test sets contained relative clauses incorporating the stem verbs *miru* (見る 'to see') and *au* (会う 'to meet'). The third test set contained relative clauses with a causative stem verb, while the fourth test set was randomly extracted from the superset of all sentences containing relative clauses[2] .

[2] The random test set used in this paper is identical to that described in Baldwin *et al.* (1997).

VERB SET: (No. sentences/clauses)		*miru* (303/308)		*au* (72/72)		CAUSATIVE (101/101)		RANDOM (129/147)	
		Orig	Rev	Orig	Rev	Orig	Rev	Orig	Rev
Overall accuracy (*No. correct*)		73.6% (226)	93.2% (287)	90.3% (65)	97.2% (70)	83.3% (85)	92.2% (93)	89.1% (131)	94.6% (139)
IDIOM	#	26		0		0		1	
	$\mathcal{P}$	*N/A*	100%	*N/A*	*N/A*	*N/A*	*N/A*	*N/A*	100%
	$\mathcal{R}$	*N/A*	100%	*N/A*	*N/A*	*N/A*	*N/A*	0%	100%
Time-related (RELT,TEMP)	#	10		6		2		5	
	$\mathcal{P}$	100%	100%	100%	100%	100%	100%	100%	100%
	$\mathcal{R}$	100%	100%	100%	100%	100%	100%	100%	100%
NOM	#	154		4		70		78	
	$\mathcal{P}$	80.9%	92.1%	*N/A*	100%	95.2%	90.8%	92.6%	92.9%
	$\mathcal{R}$	74%	98.7%	0%	100%	84.3%	98.6%	96.2%	100%
ACC	#	68		43		3		18	
	$\mathcal{P}$	57.3%	94.0%	86.0%	95.6%	13.3%	100%	85.0%	100%
	$\mathcal{R}$	98.5%	92.6%	100%	100%	66.7%	33.3%	94.4%	100%
LOC	#	3		1		0		6	
	$\mathcal{P}$	*N/A*	*N/A*	*N/A*	*N/A*	*N/A*	*N/A*	*N/A*	100%
	$\mathcal{R}$	0%	0%	0%	0%	*N/A*	*N/A*	0%	50.0%
Non-gapping clauses (NE,NON_GAPPING)	#	47		18		27		39	
	$\mathcal{P}$	89.7%	90%	100%	100%	95.7%	95.7%	82.9%	94.4%
	$\mathcal{R}$	74.5%	76.6%	88.9%	94.4%	81.5%	81.5%	87.2%	87.2%

Table 1: The accuracy of the algorithms on the given data sets

Table 1 gives the results for the original ('Orig') and revised ('Rev') algorithms on each data set, including a general accuracy and a breakdown of the performace of the algorithms on each of the principal gapping categories. The '#' row for each gapping category indicates the actual number of instances of that category, and the precision and recall are given in the $\mathcal{P}$ and $\mathcal{R}$ rows, respectively. Figures indicated as '*N/A*' were uncomputable because of a zero denominator.

As can be seen from these results, the revised algorithm outperforms the original algorithm on all data sets, and for all gapping categories. The results for the *miru* data set exemplify the benefits of the human-based rules and idiom dictionary, while the figures for the *au* data set demonstrate the validity of the concept of 'empathy-based' verbs and the treatment thereof. While there is still room for improvement in the handling of causative relative clauses, the overall error of the algorithm has been roughly halved. Initial experiments on passive clauses indicated that, here again, the overall proportion of errors was roughly halved. For the random data set, an accuracy of around 95% was achieved, suggesting this as an average performance value for the revised algorithm, as compared to around 89% for the original algorithm.

5 Conclusion

The revisions to the original algorithm proved effective, and succeeded in raising the algorithm accuracy to 95% on random data. One clear area still requiring attention is non-gapping clauses, including their further classification along semantic lines.

References

BALDWIN, T., H. TANAKA, and T. TOKUNAGA. 1997. Analysis of head gapping in Japanese relative clauses. In *IPSJ SIJ Notes*, volume 97, no. 4, 1–8.

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

IKEHARA, S., 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.

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