Making Sense of Japanese Relative Clause Constructions

Timothy Baldwin
CSLI
Stanford University
Stanford, CA 94305 USA
tbaldwin@csli.stanford.edu

Abstract

We apply the C4.5 decision tree learner in interpreting Japanese relative clause constructions, based around shallow syntactic and semantic processing. In parameterising data for use with C4.5, we propose and test various means of reducing intra-clausal interpretational ambiguity, and cross indexing the overall analysis of cosubordinated relative clause constructions. We additionally investigate the disambiguating effect of the different parameter types used, and establish upper bounds for the task.

1 Introduction

Japanese relative clause constructions have the general structure [[S][NP]], and constitute a noun phrase. We will term the modifying S the "relative clause", the modified NP the "head NP", and the overall NP a "relative clause construction" or RCC. Example RCCs are:[1]

(1) *kinō* *katta* *bōsi*
yesterday bought hat
"the hat which ( ) bought yesterday"

(2) *bōsi-o* *katta* *riyū*
hat-ACC bought reason
"the reason ( ) bought a hat"

(3) *taterareta* *yokutosi*
built next year
"the year after ( ) was built"

Different claims have been made as to the roles of syntax, semantics and pragmatics (or frame semantics) in the construal of Japanese RCCs (e.g. Teramura (1975–78), Sirai and Gunji (1998), Matsumoto (1997)). We consider two basic syntactico-semantic selection processes to govern RCC construal: selection of the relative clause by the head NP and selection of the head NP by the relative clause. These processes can be seen to be at play in the examples above: in (1), the head verb of the relative clause selects for the head NP, and a direct object case-slot gapping interpretation results (i.e. *bōsi* is the direct object of *katta*); in (2), the head NP selects for the relative clause, resulting in an attributive interpretation (i.e. *bōsi-o katta* is an attributive modifier of *riyū*); and in (3) an attributive interpretation similarly results, with the qualification that while *yokutosi* selects for the relative clause, the relative clause must in turn be able to select for a temporal modifier (e.g. stative verbs such as *soNzai-suru* "exist" are incompatible with this construction). There is a close relationship between syntax and semantics here, in that syntax provides the basic argument and modifier positions for the head verb of the relative clause, which semantics fleshes out by way of selectional restrictions. Pragmatics also has a role to play in rating the plausibility of different interpretations (Matsumoto, 1997), although we ignore its effects, and indeed the impact of context, in this research.

Our objective in this paper is, given a taxonomy of Japanese RCC semantic types (Baldwin, 1998) and a gold-standard set of Japanese RCC instances, to investigate the success of various parameter configurations in interpreting RCCs. One feature of the proposed method is that it is based on shallow analysis, centring principally around a basic case frame and verb class description. That is, we attempt to make maximum use of surface information in performing a deep semantic task, in the same vein, e.g., as Joanis and Stevenson (2003) for English verb classification and Lapata (2002) in disambiguating nominalisations.

Relative clause interpretation is a core component of text understanding, as demonstrated in the context of the MUC conference series (Cardie, 1992; Hobbs et al., 1997). It also has immediate applications in, e.g., Japanese–English machine translation: for case-slot gapping RCCs such as (1), we extrapose the head NP from the appropriate argument position in the English relative clause (producing,

[1] The following abbreviations are used in glosses: NOM = nominative, ACC = accusative, PRES = non-past and POT = potential. ( ) is used to indicate zero (anaphoric) arguments.

e.g., "the hat$_i$ [e_i bought yesterday]"), and for attributive RCCs such as (2), we generate the English relative clause without extraposition and select the relative pronoun according to the head NP (producing, e.g., "the reason that the hat was bought").

RCC interpretation is dogged by analytical ambiguity, in particular for phrase boundary, phrase head/attachment and word sense ambiguity. The first two of these concerns can be dealt with by a parser such as KNP (Kurohashi and Nagao, 1998) or CaboCha (Kudo and Matsumoto, 2002), or alternatively a tag sequence-based technique such as that proposed by Siddharthan (2002) for English. Word sense ambiguity is an issue if we wish to determine the valence of the verb and make use of selectional restrictions. We sidestep full-on verb sense disambiguation by associating a unique case frame with each verb stem type and encoding common alternations in the verb class. Even here, however, we must have some means of dealing with verb homonymy and integrating analyses for cosubordinated relative clauses. We investigate various techniques to resolve such ambiguity and combine the analysis of multiple component clauses.

In the following, we define the RCC semantic types (§ 2) and outline the parameters used in the proposed method (§ 3). We then discuss sources of ambiguity and disambiguation methods (§ 4), before evaluating the proposed methods (§ 5), and finally comparing the results with those of previous research (§ 6).

2 Definitions

We define relative clause modification as falling into three major semantic categories, indistinguishable orthographically: case-slot gapping, attributive and idiomatic.

Case-slot gapping RCCs (aka "internal"/"inner relation" (Teramura, 1975–78) or "clause host" RCCs (Matsumoto, 1997)), are characterised by the head NP having been gapped (or extraposed) from a case slot subcategorised by the main verb of the relative clause (see (1)). For our purposes, case-slot gapping is considered to occur in 19 sub-categories, which can be partitioned into 8 **argument case slot** types (e.g. SUBJECT, DIRECT OBJECT, INDIRECT OBJECT) and 11 **modifier case slot** types (e.g. INSTRUMENT, TEMPORAL, SOURCE LOCATIVE: Baldwin (1998)). Note that the case marking on the slot from which gapping has occurred is not preserved either within the relative clause or on the head NP.

Attributive RCCs (aka "external"/"outer relation" (Teramura, 1975–78) or "noun host" RCCs (Matsumoto, 1997)) occur when the relative clause modifies or restricts the denotatum of the head NP (see (2)). They come in 7 varieties according to the nature of modification (e.g. CONTENT, RESULTATIVE, EXCLUSIVE).

Idiomatic RCCs are produced when the overall RCC produces a constructionally idiomatic reading, e.g.:

(4) *mite* *minu* *huri*
to see not see pretend
"looking the other way"

One feature of idiomatic RCCs is that they can be described by a largely lexicalised construction template, and are incompatible with conjugational alternation and modifier case slots. Due to the non-compositional nature of idiomatic RCCs, we make no attempt to analyse them by way of the case-slot gapping/attributive RCC dichotomy, or sub-classify them further.

Japanese RCC interpretation as defined in this paper is according to the 27 interpretation types subsumed by these 3 basic categories of RCC construal. It is important to realise that these interpretation types are lexically indistinguishable. The semantic type of the RCC is therefore not readily accessible from a simple structural analysis of the RCC as contained within a standard treebank.

3 Parameter description

Features used in the interpretation of RCCs include a generalised case frame description, a verb class characterisation, head noun semantics, morphological analysis of the head verb, and various constructional templates. These combine to form the 49-feature parameter signature of each RCC. Unless otherwise mentioned, all features are binary.

Case frames are applied in determining which argument case slots are subcategorised by the head verb of the relative clause and instantiated—hence making them *un*available for case-slot gapping—and conversely which case slots are subcategorised by the head verb and *un*instantiated—making them available for case slot gapping. The range of argument case slots coincides exactly with the set of argument case-slot gapping RCC types from § 2 (8 features in total).

Argument case slot instantiation features are set by comparing a given case frame to the actual input, and aligning case slots between the two according to case marker correspondence. In the case frame dictionary, a single generalised case frame is given for each verb stem. Case frames were generated

from the Goi-Taikei pattern-based valency dictionary (Ikehara et al., 1997) by manually merging the major senses for each distinct verb stem. In essence, case frames are simply a list of the argument case slots for the verb in question in their canonical ordering (case frames include no modifier case slots). Each case slot is marked for canonical case marking and case slot type.

Case frames can contain lexicalised case slots, which must be overtly realised for that case frame to be triggered. Examples of fixed expressions are *ki-o tukeru* (mind-ACC fix/attach) "to be careful/keep an eye out for (something)" and *yume-o miru* (dream-ACC see) "to dream". We manually annotated each fixed argument for "gapability", i.e. the potential for extraposition to the head NP position such as with the RCC *kinō mita yume* "the dream I had last night". If a gapable fixed argument occurs (unmodified) in head NP position, we use the "gapped fixed argument head NP" feature to return the argument type of gapped fixed argument (e.g. DIRECT OBJECT).

The unique case frame description is complemented by **verb classes**. Verb classes are used to describe such effects as: (1) modifier case slot compatibility, e.g. PROXIMAL verbs such as *kaeru* "return" are compatible with target locative modifier case slots; (2) case slot interaction, e.g. INTERPERSONAL verbs such as *au* "meet" have two coindexed argument slots to indicate the interacting parties; and (3) potential for valency-modifying alternation, e.g. INCHOATIVE verbs such as *kaisi-suru* "start" are listed with the (unaccusative) intansitive case frame but undergo the causative-inchoative alternation to produce transitive case frames (Jacobsen, 1992). A total of 27 verb classes are used in this research, which incorporate a subset of the verbal semantic attributes (VSAs) of Nakaiwa and Ikehara (1997) as well as classes independently developed for the purposes of this research.

Head noun semantics are used to morpho-semantically classify the head noun (of the head NP) into 14 classes (e.g. AGENTIVE, TEMPORAL, FIRST-PERSON PRONOUN), based on the Goi-Taikei noun taxonomy. Rather than attempting to disambiguate noun sense, the head noun semantic features are determined as the union of all senses of the head noun of the head NP. For coordinated head NPs, we take the intersection of the head noun feature vectors. One head noun semantic feature particular to RCCs is the class of functional nouns (e.g. *riyū* "reason", *kekka* "result" and *mokuteki* "objective") which generally give rise to attributive RCCs.

In processing each unit relative clause, we carry out **morphological analysis** of the head verb of the relative clause, returning a listing of verb morphemes and tense/aspect affixes: e.g. the verb *okonawareteita* "to have been held" is analysed as *okona-ware-te-ita* "to hold-PASSIVE-PROGRESSIVE-PAST". This has applications in case frame transformation (e.g. passivisation), as trigger conditions in constructional templates, and in the resolution of case frame ambiguity. Case frame transformation is carried out prior to matching case slots between the input and case frame, producing a description of the surface realisation of the case frame which reflects the voice, causality, etc. of the main verb. Case frame transformation can potentially produce fan-out in the number of clause analyses, particularly in the case of the *(r)are* verb morpheme, which has passive, potential/spontaneous and honorific readings (Jacobsen, 1992). We produce all legal case frames in this case, and leave the selection of the correct verb interpretation for later processing. Note that the only morphological verb feature to make an appearance as an independent feature is POTENTIALITY, as it combines with nominalised adjectives to produce COMPARATIVE RCCs such as *tob-eru hirosa* (jump-POT size) "(of) size big enough to jump (in)".

In addition to simple features, there are a number of **constructional templates**, namely two features for the attributive RCC types of EXCLUSIVE and INCLUSIVE, and also one feature for idiomatic RCCs. The constructional template for EXCLUSIVE RCCs operates over the EXCLUDING verb class (containing *nozoku* "to exclude", for example), and stipulates simple past or non-past main verb conugation and the occurrence of only an accusatively-marked case slot within the relative clause. The satisfaction of these constraints results in the EXCLUSIVE RCC compatibility feature being set, as occurs for:

(5) *nitiyōbi-o* *nozo-ku* *mainiti*
Sunday-ACC exclude-PRES everyday
"every day except Sundays"

Idiomatic RCC templates constrain the lexical type and modifiability of the head NP, verbal conjugation, case marker alternation and modifier case slots/adverbials. A total of 11 templates are utilised in the current system, which are mapped onto a single feature value.

4 Analytical ambiguity and disambiguation

As with any NLP task, ambiguity occurs at various levels in the data. In this section, we outline sources

of ambiguity and propose disambiguation methods for each.

4.1 Analytical ambiguity

Analytical ambiguity arises when multiple clause analyses exist, as a result of verb homophony/homography or fixed expression compatibility.

For the purposes of our system, **verb homophony** occurs when multiple verb entries in the case frame dictionary share the same kana content (and hence pronunciation), such that a kana-based orthography will lead to ambiguity between the different entries. **Verb homography**, on the other hand, occurs when multiple verb entries coincide in kanji content, leading to ambiguity for a kanji-based orthography. Both verb homophony and homography can be either full or partial, i.e. all forms of a given verb pair can be homophonous/homographic, or there can be partial overlap for particular types of verb inflection. For example, the verbs 変わる *kawaru* "change" and 代わる *kawaru* "replace" are fully homophonous, whereas 着る *kiru* "wear" and 切る *kiru* "cut" are partially homophonous (e.g., in the simple non-past they diverge in kana orthography, producing *kita* and *kitta*, respectively). For verb homography, 止める *tomeru* "stop" and 止める *yameru* "quit" are fully homographic, whereas 行う *okonau* "carry out" and 行く *iku* "go" are partially homographic (with overlap produced for the simple past tense, e.g., in the form of 行った, which can be read as either *okonatta* or *itta*). Such overlap in lexical form leads to the situation of multiple verb entries being triggered, producing independent analyses for the RCC input.

Fixed expressions lead to analytical ambiguity as, in most cases, the main verb of the expression will also be compatible with productive usages, by way of a generalised case frame entry. For example, in addition to the fixed expression *asi-o arau* (foot-ACC wash) "quit", *arau* "wash" has a (unique) non-lexicalised case frame entry, which will be compatible with any lexical context satisfying the lexical constraints on the fixed expression.

4.2 Resolving analytical ambiguity

Here, we present a cascaded system of heuristics which resolves analytical ambiguity arising from multiple verb entries, producing a unique feature vector characterisation.

We select between multiple analyses for a given relative clause in the first by preferring analyses stemming from fixed expressions, over those conforming to constructional templates, in turn over those generated through generalised techniques. We define each such stratum as comprising a distinct **expressional type**, similarly to Ikehara et al. (1996).

Expressional type is on the whole a simple but powerful disambiguation mechanism, but is not infallible. The main area in which it comes unstuck is in giving fixed expressions absolute priority over other analyses. Many fixed expressions can also be interpreted compositionally: e.g. *asi-o arau* (foot-ACC wash) "quit" can mean simply "wash (one's) feet". In the case of *asi-o arau*, the case frame is identical between the fixed and generalised expression, but the verb classes are significanly different, potentially leading to unfortunate side-effects when trying to interpret an RCC involving the non-idiomatic sense of the verb.

Fixed expressions and RCCs compatible with constructional templates tend to be relatively rare, so in most cases, ambiguity is not resolved through expressional type preferences. In this case, we apply a succession of heuristics of decreasing reliability, until we produce a unique analysis and feature vector characterisation. These heuristics are, in order of application: minimum verb morpheme content, best case frame match and representational preference.

Minimum verb morpheme content involves determining the morphemic content of the head verb of the relative clause for each verb stem it is compatible with, and selecting the verb stem(s) which are morphologically least complex. Morphological complexity is determined by simply counting the number of morphemes, auxiliary verbs and affixes in the verb composite. Given the verb composite 見える *mieru* e.g., we would generate two analyses: *mie-ru* "can see-PRES" and *mi-e-ru* "see-POT-PRES", of which we would (correctly) select the first. In essence, this methodology picks up on more highly stem-lexicalised verb entries, and effectively blocks more compositional verb entries.

With **best case frame match**, we analyse the degree of correspondence between the case frame listed for each dictionary entry, and the actual case slot content of the input. In following with the shallow processing objective of this research, we simply calculate the number of case slots in the input which align with case slots in each case frame (based on case marker overlap), and divide this by the sum of the case slots in the case frame and in the input. We additionally add one to the numerator to give preference to case frames of lower valency (i.e. fewer case slots) in the case that there is no overlap with

the input. This can be formalised as:

$$CFM(IN, CF) \quad = \quad \frac{1 + |IN \hat{\cap} CF|}{|IN| + |CF|}$$

where IN is the set of case slots in the input, CF the set of case slots in the current case frame, and $\hat{\cap}$ the case slot overlap operator. Note that the ordering of the case slots plays no part in calculations, in an attempt to capture the relative freedom of case slot order in Japanese.

The final heuristic is of high recall but lesser precision, to resolve any remaining ambiguity. It is based on the **representational preference** for the current verb to take different lexical forms. The representational preference (RP) of lexical form a of verb entry f (i.e. a_f) is defined as the likelihood of f being realised as a:

$$RP(a_f) \quad = \quad \frac{1 + freq(a_f)}{1 + \sum_{i \neq a} freq(i_f)}$$

This is normalised over the representational preference for all source entries a_i, producing the verb score (VS) for each a_f:

$$VS(a_f) \quad = \quad \frac{RP(a_f)}{\sum_i RP(a_i)}$$

All frequencies are calculated based on the EDR corpus (EDR, 1995), a 2m morpheme corpus of largely technical Japanese prose.

In the case of a tie in representational preference, we select one of the tied analyses randomly.

4.3 Clause cosubordination and disambiguation

Japanese cosubordinated clauses (i.e. dependent but not embedded clauses, as indicated by the use of a conjunction such as *nagara*, *te*, *tutu* or *si*, or through *continuative* type conjugation: Van Valin (1984)) offer an additional avenue for disambiguation:

(6) *[[Kim-ga* Kim-NOM *kōaN-si,* design *] seisaku-sita* produced *] kikai*
machine
"a machine designed and produced by Kim"

(7) *[[kyoneN* last year *hatumei-sare* invented *] ryūkō-sita* got popular *] mono*
thing
"things which were invented and gained popularity last year"

As is apparent in (6) and (7), a consistent RCC interpretation is maintained across cosubordinated clauses, e.g. in (6), *kikai* "machine" is the DIRECT OBJECT of both *kōaN-si* and *seisaku-sita*.[2] It is possible to put this observation to use when interpreting cosubordinated RCCs, by coordinating the feature vectors for the unit clauses to produce a unique, coherent interpretation for the overall RCC. We apply this in two ways: by OR'ing and AND'ing the feature vectors together.

5 Evaluation

In evaluation, we compare different clausal interpretation selection techniques. We further go on to investigate the efficacy of different parameter partitions on disambiguation, and generate a learning curve.

Evaluation was carried out by way of stratified 10-fold cross validation throughout, using the C4.5 decision tree learner (Quinlan, 1993).[3] As C4.5 induces a unique decision tree from the training data and then applies this to the test data, we are able to evaluate both training and test classification accuracy, i.e. the relative success of the decision tree in classifying the training data and test data, respectively.

The data used in evaluation is a set of 5143 RCC instances from the EDR corpus (EDR, 1995), of which 4.7% included cosubordinated relative clauses (i.e. the total number of unit relative clauses is 5408). Each RCC instance was manually annotated for default interpretation independent of sentential context. The 10 most-frequent interpretations (out of 27) in this test set are presented below:

Interpretation	***RCC supertype***	***Freq***
SUBJECT	case-slot gapping	.640
CONTENT	attributive	.135
DIRECT OBJECT	case-slot gapping	.074
IDIOMATIC	idiomatic	.024
EXCLUSIVE	attributive	.023
LOCATIVE	case-slot gapping	.022
TEMPORAL	case-slot gapping	.021
CO-SUBJECT	case-slot gapping	.012
STATIVE TOPIC	case-slot gapping	.010
TIME DURATIONAL	case-slot gapping	.009

Based on this, we can derive a baseline accuracy of 64.0%, obtained by allocating the SUBJECT interpretation to every RCC input.

[2] Note that in (7), the SUBJECT interpretation is shared between a passive and active clause. It is because the interpretational parallelism occurs at the grammatical relation level rather than case-role level that we select grammatical relations for our argument case-slot gapping types.

[3] We also ran TiMBL 5.0, TinySVM and Rob Malouf's MaxEnt toolkit over the data, but found C4.5 to produce the best results.

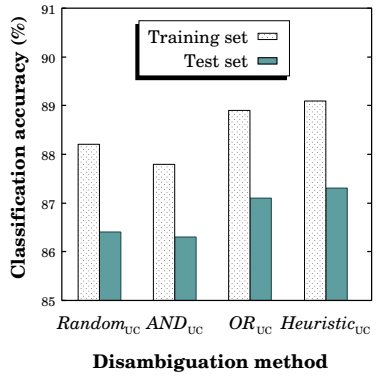

Figure 1: Evaluation of unit clause disambiguation strategies

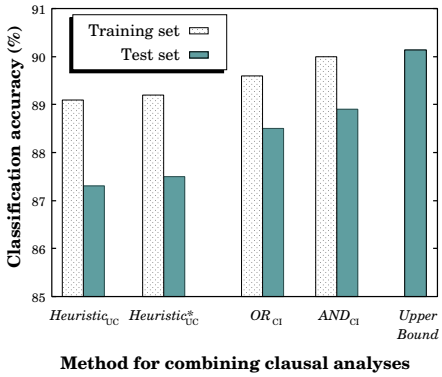

Figure 2: Evaluation of cosubordinated clause disambiguation strategies

5.1 Evaluation of analytical disambiguation

First, we evaluate analytical disambiguation by decomposing each RCC into its component cosubordinated RCCs and selecting most plausible interpretation for each unit clause (*UC*). We compare: (a) a random selection baseline method ($Random_{UC}$); (b) a method where all feature vectors for the unit relative clause are logically AND'ed together (AND_{UC}); (c) a method where all feature vectors for the unit clause are logically OR'ed together (OR_{UC}); and (d) the cascaded-heuristic method from § 4.2 above ($Heuristic_{UC}$). The results for the various methods are presented in Fig. 1. Note that 28.8% of clauses occurring in the data are associated with analytical ambiguity, and for the remainder, there is only one verb entry in the case frame dictionary.

$Heuristic_{UC}$ outperforms the $Random_{UC}$ baseline to a level of statistical significance,[4] in both training and testing. OR_{UC} lags behind $Heuristic_{UC}$ in testing in particular, but is vastly superior to AND_{UC}, which is marginally worse than $Random_{UC}$ in both training and testing.

Based on these results, we conclude that our system of cascaded heuristics ($Heuristic_{UC}$) is the best of the tested methods and use this as our intra-clause disambiguation method in subsequent evaluation.

[4] All statistical significance judgements are based on the paired t test ($p < 0.05$).

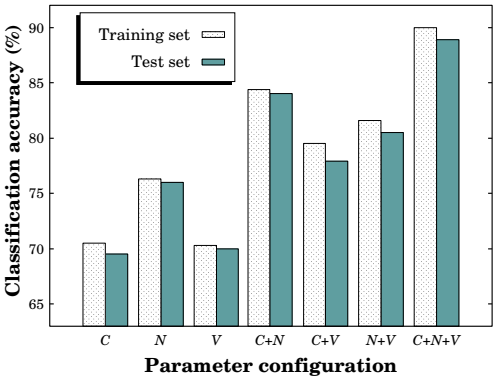

Figure 3: Evaluation of different parameter combinations (*C* = case slot instantiation, *N* = head noun semantics, and *V* = head verb class)

5.2 Disambiguation via cosubordination

Next, we test the cosubordination-based disambiguation techniques. The two core paradigms we consider are: (1) unit clause (*UC*) analysis, where each cosubordinated clause is considered independently, as in § 5.1; and (2) clause-integrated (*CI*) analysis, where we actively use cosubordination in disambiguation.

For **unit clause analysis**, we replicate the basic $Heuristic_{UC}$ methodology from above and also extend it by logically AND'ing together the case slot instantiation flags between unit clause feature vectors to maintain a consistently applicable case-role gapping analysis ($Heuristic^{*}_{UC}$).

For **clause-integrated analysis**, we apply *Heuristic* in intra-clausal analysis, then either logically OR or AND the component unit clause feature vectors together, producing methods OR_{CI} and AND_{CI}, respectively.

The training and test accuracies for the described methods over the full data set are given in Fig. 2.

$Heuristic^{*}_{UC}$ (incorporating inter-clausal coordination of only case slot data) appears to offer a slight advantage over $Heuristic_{UC}$, but the two clause-integrated analysis methods of OR_{CI} and AND_{CI} are significantly superior in both testing and training. Overall, the best-performing method is AND_{CI} at a test accuracy of 88.9%.

It is difficult to gauge the significance of the results given that coordinating RCC's account for only 4.7% of the total data. One reference point

is the performance of the *Heuristic*$_{\text{UC}}$ method over only simple (non-cosubordinated) RCCs. This gives a training accuracy of 90.6% and test accuracy of 89.3%, suggesting that we are actually doing slightly worse over cosubordinated RCCs than simple RCCs, but that we gain considerably from employing a clause-integrated approach relative to simple unit clause analysis.

An absolute cap on performance for the original system can be obtained through *non-deterministic* evaluation, whereby the system is adjudged to be correct in the instance that the correct analysis is produced for any one unit clause analysis (out of the multiple analyses per clause). This produces an accuracy of 90.2%, which is presented as *Upper Bound* in Fig. 2. Given that all that the proposed method is doing is choosing between the different unit clause analyses, it cannot hope to better this. Relative to the baseline and upper bound, the error reduction for the clause-integrated *AND*$_{\text{CI}}$ method is 96.6%, a very strong result.

5.3 Additional evaluation

We further partitioned up the parameter space and ran C4.5 over the different combinations thereof, using *AND*$_{\text{CI}}$. The particular parameter partitions we target are case slot instantiation flags (*C*: 11 features), head noun semantics (*N*: 14 features) and verb classes (*V*: 27 features).

The system results over the individual parameter partitions, and the various combinations of case slot instantiation, head noun semantics and verb classes (e.g. *N+V* = head noun semantics and verb classes), are presented in Fig. 3.[5]

The value of head noun semantics is borne out by the high test accuracy for *N* of 76.0%. We can additionally see that case slot instantiation and verb class features provide approximately equivalent discriminatory power, both well above the absolute baseline of 64.0%. This is despite case slot instantiation flags being less than half the number of verb classes, largely due to the direct correlation between case slot instantiation judgements and case-slot gapping analyses, which account for around 80% of all RCCs.

The affinity between case slot instantiation judgements and the semantics of the head noun is evidenced in the strong performance of *C+N*, although even here, verb classes gain us an additional 5% of performance. Essentially what is occurring here is that selectional preferences between particular head noun semantics and certain case-slot/analysis types are incrementally enhanced as we add in the extra dimensions of case slot instantiation and verb classes. The orthogonality of the three dimensions is demonstrated by the incremental performance improvement as we add in extra parameter types. This finding provides evidence for our earlier claims about selection in RCCs being based on the combination of head noun semantics, verb classes and information about what case slots are vacant in the relative clause.

To determine if the 90.2% upper bound on classification accuracy for the given experimental setup is due to limitations in the particular resources we are using or an inherent bound on the RCC interpretation task as defined herein, we performed a manual annotation task involving 4 annotators and 100 randomly-selected RCCs, taken from the 5143 RCCs used in this research. The mean agreement between the annotators was 90.0%, coinciding remarkably well with the 90.2% figure. This provides extra evidence for the success of the proposed method, and suggests that there is little room for improvement given the current task definition.

6 Discussion

Perhaps the most directly comparable research to that outlined in this paper is that of Abekawa et al. (2001), who disambiguate RCCs according to simplex dependency data and KL divergence. That is, they extract out $(noun, case\_marker, verb)$ triples from corpus data, and disambiguate RCCs according to which case slot the head noun occurs in most commonly in simplex data. The accuracy for their method over a task where they distinguished between attributive and 6 types of case-slot gapping RCCs (defined according to case marker) was a relatively modest 65.3%. For a binary attributive vs. case-slot gapping task, the accuracy was a more respectable 88.8%, but still considerably lower than that achieved in this research.

An alternate point of reference is found in the work of Li et al. (1998) on Korean RCCs, which display the same structural ambiguities as Japanese RCCs. Li et al. (1998) attain an accuracy of 90.4% through statistical analysis of the distribution of verb-case filler collocates, except that they classify relative clauses according to only 5 categories and consider only case-slot gapping RCCs. With our method, restricting analysis to only gapping RCCs (still retaining a total of nineteen RCC types) produces an accuracy of 94.1% for the *AND*$_{\text{CI}}$ system with C4.5.

In conclusion, we have proposed a method for in-

[5] Note that *C+N+V* corresponds to the full parameter space, and is identical to *AND*$_{\text{CI}}$ in Figure 2.

terpreting Japanese relative clause constructions according to surface evidence and a generalised semantic representation. The method is designed to cope with analytical ambiguity in the head verb and head noun, and also interpretational parallelism in cosubordinated RCCs. In evaluation using C4.5, we showed our system to have a classification accuracy of 89.3%, marginally below the 90% upper bound for the described task.

We have totally ignored the effects of pragmatics and context in this research, and in doing so, shown that it is possible to reliably derive a default RCC interpretation using only shallow syntactic and semantic features. In future research, we are interested in exploring methods of incorporating pragmatic and contextual features into our method, and the impact of these factors on both human and machine RCC interpretation.

Acknowledgements

This material is based upon work supported by the National Science Foundation under Grant No. BCS-0094638 and was partially conducted while the author was an invited researcher at the NTT Communication Science Laboratories, Nippon Telegraph and Telephone Corporation. We would like to thank Emily Bender, Francis Bond, Kenji Kimura, Christoph Neumann, Tomoya Noro, Satoko Shiga, Hozumi Tanaka and the various anonymous reviewers for their valuable input on this research.

References

Takeshi Abekawa, Kiyoaki Shirai, Hozumi Tanaka, and Takenobu Tokunaga. 2001. *Tōkei-jōhō-o riyō-shita Nihongo-rentai-shūshoku-setsu no kaiseki* (statistical analysis of Japanese relative clause constructions). In *Proc. of the 7th Annual Meeting of the Association for Natural Language Processing (Japan)*, pages 269–72, Tokyo, Japan. (in Japanese).

Timothy Baldwin. 1998. *The Analysis of Japanese Relative Clauses*. Master's thesis, Tokyo Institute of Technology.

Claire Cardie. 1992. Corpus-based acquisition of relative pronoun disambiguation heuristics. In *Proc. of the 30th Annual Meeting of the ACL*, pages 216–23, Newark, USA.

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

Jerry R. Hobbs, Douglas Appelt, John Bear, David Israel, Megumi Kameyama, Mark Stickel, and Mabry Tyson. 1997. FASTUS: A cascaded finite-state transducer for extracting information from natural-language text. In Emmanuel Roche and Yves Schabes, editors, *Finite State Devices for Natural Language Processing*. MIT Press, Cambridge, USA.

Satoru Ikehara, Satoshi Shirai, and Francis Bond. 1996. Approaches to disambiguation in ALT-J/E. In *Proc. of the International Seminar on Multimodal Interactive Disambiguation: MIDDIM-96*, pages 107–17, Grenoble, France.

Satoru Ikehara, Masahiro Miyazaki, Akio Yokoo, Satoshi Shirai, Hiromi Nakaiwa, Kentaro Ogura, Yoshifumi Ooyama, and Yoshihiko Hayashi. 1997. *Nihongo Goi Taikei – A Japanese Lexicon*. Iwanami Shoten. 5 volumes. (In Japanese).

Wesley M. Jacobsen. 1992. *The Transitive Structure of Events in Japanese*. Kurosio Publishers.

Eric Joanis and Suzanne Stevenson. 2003. A general feature space for automatic verb classification. In *Proc. of the 10th Conference of the EACL (EACL 2003)*, pages 163–70, Budapest, Hungary.

Taku Kudo and Yuji Matsumoto. 2002. Japanese dependency analysis using cascaded chunking. In *Proc. of the 6th Conference on Natural Language Learning (CoNLL-2002)*, pages 63–9, Taipei, Taiwan.

Sadao Kurohashi and Makoto Nagao. 1998. Building a Japanese parsed corpus while improving the parsing system. In *Proc. of the 1st International Conference on Language Resources and Evaluation (LREC'98)*, pages 719–24.

Maria Lapata. 2002. The disambiguation of nominalizations. *Computational Linguistics*, 28(3):357–88.

Hui-Feng Li, Jong-Hyeok Lee, and Geunbae Lee. 1998. Identifying syntactic role of antecedent in Korean relative clause using corpus and thesaurus information. In *Proc. of the 36th Annual Meeting of the ACL and 17th International Conference on Computational Linguistics (COLING/ACL-98)*, pages 756–62, Montreal, Canada.

Yoshiko Matsumoto. 1997. *Noun Modifying Constructions in Japanese*. John Benjamins.

Hiromi Nakaiwa and Satoru Ikehara. 1997. A system of verbal semantic attributes in Japanese focused on syntactic correspondence between Japanese and English. *Journal of the Information Processing Society of Japan*, 38(2):215–25. (In Japanese).

J. Ross Quinlan. 1993. *C4.5: Programs for Machine Learning*. Morgan Kaufmann.

Advaith Siddharthan. 2002. Resolving attachment and clause boundary ambiguities for simplifying relative clause constructs. In *Proc. of the Student Research Workshop, 40th Annual Meeting of the ACL (ACL-02)*, pages 60–5, Philadelphia, USA.

Hidetosi Sirai and Takao Gunji. 1998. Relative clauses and adnominal clauses. In Takao Gunji and Koiti Hasida, editors, *Topics in Constraint-Based Grammar of Japanese*, chapter 2, pages 17–38. Kluwer Academic, Dordrecht, Netherlands.

Hideo Teramura. 1975–78. Rentai-shushoku no shintakusu to imi Nos. 1–4. In *Nihongo Nihonbunka 4–7*, pages 71–119, 29–78, 1–35, 1–24. Osaka: Osaka Gaikokugo Daigaku. (In Japanese).

Robert Van Valin. 1984. A typology of syntactic relations in clause linkage. In *Proc. of the Tenth Annual Meeting of the Berkeley Linguistics Society*, pages 542–58.