Disambiguating Japanese Compound Verbs

Kiyoko Uchiyama\textsuperscript{a,b,*}, Timothy Baldwin\textsuperscript{b} Shun Ishizaki\textsuperscript{a}

\textsuperscript{a}Keio University
5322 Endo, Fujisawa-shi, Kanagawa 252-8520 JAPAN
\textsuperscript{b}Center for the Study of Language and Information
Stanford University
210 Panama street, Stanford, CA 94305 USA

Abstract

The purpose of this study is to disambiguate Japanese compound verbs (JCV) based on two methods: (1) a statistical method which makes use of collocational or semantic information about different verb combinations, and (2) a manual rule-based method which utilises verbal and nominal semantic features. We also present a combined method where the output of the statistical method is fed into the rule-based method. In evaluation, we found that the pure rule-based method outperformed the statistical and combined method at 96\% token-level accuracy, suggesting that fine-grained semantic analysis is an important component of JCV disambiguation. At the same time, the performance of the fully-automated statistical method was found to be surprisingly good at 86\%, without making use of lexical semantics.

Key words: disambiguation, Japanese compound verb, support vector machine, verb semantics

1 Introduction

Multiword expressions (MWEs) have been identified as a primary bottleneck in parsing and language understanding (Sag et al., 2002; Baldwin and Bond, 2002). For the purposes of this paper, we follow Baldwin and Bond (2002) in defining MWEs to be “idiosyncratic interpretations that cross word boundaries”, and focus on the issue of idiomaticity in the context of Japanese compound verbs. Idiomaticity is a statement of the difficulty in predicting the

* Corresponding author.

Email addresses: kiyoko@sfc.keio.ac.jp, tbaldwin@csli.stanford.edu, ishizaki@sfc.keio.ac.jp (Shun Ishizaki).

Preprint submitted to Elsevier Science 7 June 2004
semantics of a given MWE, e.g. in predicting that a group photograph is a photograph depicting a group whereas a colour photograph is a photograph in colour. We target the semi-productive Japanese compound verb (JCV) construction in illustrating the nature of idiomaticity and presenting a method to deal with its effects. JCVs represent the concatenation of two verbs, the first of which always appears in the continuative form, as inoshi-ageru “push up” and tabe-sugiru “eat too much”;¹ throughout this paper, we will refer to the first verb as the V1 and the second verb as the V2. For the purposes of this research, we assume that both the V1 and V2 are native Japanese verbs, thus excluding compounds such as teNtô-sidasu “start to lean over”. JCVs are frequently used to express directed motions (e.g. uki-agaru “float up”), elaborated phenomena (e.g. hare-wataru “clear up”) and emotional states (e.g. kanji-iru “be deeply impressed”). They are characteristically highly productive and semantically ambiguous, and are subject to semantic constraints between the V1 and V2. Examples of JCV semantic constraints and ambiguities are given in (1).

(1) a. nage-ageru “throw up”, keri-ageru “kick up”, moti-ageru “lift up”  
b. yude-ageru “finish boiling”, musi-ageru “finish steaming”  
yaki-ageru “finish baking”

Ageru “raise” has multiple meanings when it appears in the V2 position of a JCV. In (1-a), directional compound verbs are formed by combining the V2 ageru “raise” with a V1 verb of motion like nageru “throw” or keri “kick”. Conversely, in (1-b), an aspectual compound verb is formed by combining the V2 ageru “raise” with a V1 cooking verb like yuderu “boil” and musu “steam”. There is a small number of verbs which can occur as the V2, but they tend to combine relatively freely with a wide range of V1s, and be interpreted differently according to the semantic properties of the V1. Ambiguous JCVs are generated by compounding different V1s with an ambiguous V2, presenting problems for automated language understanding tasks.

JCVs have received a moderate amount of attention in the fields of linguistics and natural language processing. In linguistics, JCVs have been studied mainly in terms of syntax (Kageyama, 1999) and constraints on semantic structures (Matsumoto, 1996, 1998). Himeno (2001) performed a semantic analysis concerning the types of V2s which have multiple meanings. She classified JCVs according to the meaning of the V2, but in the process, blurred the boundary between V1 and V2 semantics. In order to clarify the semantic constraints between V1 and V2, we claim that it is necessary to analyse the JCV compositionally based on the individual semantics of the V1 and V2.

¹ The following abbreviations are used in glosses in this paper: TOP = topic, ACC = accusative and DAT = dative.
Shirai et al. (1998) proposed a computational method for compiling a database of JCV valency patterns based on Japanese and English corpora. Their approach was shown to improve the performance of a machine translation (MT) system. However, it is inefficient and impractical to expect to be able to predict all V1-V2 combinations in advance and precompile a lexicon of JCVs based thereupon. For that reason, we suggest that it is desirable to develop a robust framework which is able to dynamically analyse JCVs.

We propose a disambiguation method which identifies the meaning of JCVs using two basic steps: (1) a statistical approach which learns the basic senses of arbitrary V1-V2 combinations independent of context, and (2) a rule-based approach which utilises semantic features and syntactic information derived from the context of use of a given JCV token.

In the statistical approach, we try to identify the range of meanings a given JCV type can take. A significant number of JCVs are monosemous (have a unique sense), and thus do not require token-level disambiguation based on their context of use. Polysemous JCVs, on the other hand, often do not occur with the full range of interpretations possible for a JCV. In both these cases, the identification of type-level sense provides a valuable sense determinant/filter. We identify JCV semantics using collocation information extracted automatically from a corpus, and a support vector machine classifier.

The rule-based approach is run either as a standalone method or only over those JCVs that the statistical approach identifies as ambiguous. It is made up of two steps: (1) identify the semantics of the V1 using a lexical database, and (2) classify JCVs into classes based on the semantics of the V1 (semantic information) and contextual information, particularly verb complements (syntactic information). The rules were hand built based on token instances of JCVs involving a given V2.

The proposed combined method has the advantage of being able to analyse novel V1-V2 combinations. Additionally, in combining the statistical method with the rule-based method, we are able to minimise the use of lexical semantic information and external resources by identifying and filtering off monosemous JCVs.

This research has applications in language understanding tasks and MT, e.g. Japanese-to-English. One key point of interest is the correlation between Japanese compound verbs and English verb particle constructions (VPCs: Baldwin and Villavicencio (2002); Bannard et al. (2003); Villavicencio and Copestake (2002); Villavicencio (2003)). We claim that verb particles in English and compound verbs in Japanese have commonalities in terms of their ambiguity and semantic constraints, and are interested in exploiting these in an MT context. For example, the English particle up has both an aspec-
tual (write up, cf. kaki-ageru) and a spatial meaning (kick up the ball, cf. keri-ageru), which is equivalent to the V2 in Japanese compound verbs (with ageru corresponding to up in this case). Our formulation of semantic classes is designed to reflect this semantic parallelism and make our method amenable with crosslingual applications.

The remainder of the paper is structured as follows. Section 2 defines and outlines the semantic nature of JCVs. We then propose and evaluate a statistical filter for determining JCV sense at the type level in Section 3, and a rule-based method for determining JCV sense at the token level in Section 4.

2 The Semantic Nature of JCVs

2.1 JCV ambiguity

Kageyama (1993) has proposed that JCVs can be analysed according to the argument structure of each constituent and divided into two basic types: lexical and syntactic compounds. Lexical compounds (e.g. yude-ageru “boil up”) are limited to lexically-specified combinations and the V2 undergoes regular semantic alternation (e.g. ageru “raise” taking on completive semantics, derived from the simplex semantics), whereas syntactic compounds (e.g. kaki-wasureru “forget to write”) are fully productive and semantically compositional. We do not differentiate between these two types as our semantic classification allows us to deal with the semantics of both types within a single framework, and the syntactic distinction between them is irrelevant for the purposes of this research. There are two types of JCV ambiguity: ambiguities within lexical compounds (e.g. kake-agaru “run up” (spatial) and furue-agaru “be terrified”) (adverbial), and ambiguities between lexical compounds (e.g. tōri-sugiru “pass through”) and syntactic compounds (e.g. tabe-sugiru “eat too much”).

In this research, we focus on lexical compounds which contain an ambiguous V2 (as in (1) above). Semantic constraints govern the V1-V2 compounding process, such that while yude-ageru “boil up” is a legal JCV, *sini-ageru “die up (intended)” is not. The semantic properties of the V1 play a key role in determining the meaning of the V2. We focus on extracting commonalities in the semantic properties of the V1s that combine with a given V2 in order to disambiguate the sense of the V2. On the other hand, some JCVs are ambiguous between a syntactic and lexical interpretation, and can only be disambiguated given context, in which syntactic information plays an important role:

(2)  a. Basu wa basutei o hasiri-sugita.
    The bus drove past the bus stop.
b. Kare wa saia no tame ni hasiri-sugita.
He ran too much due to the game.

In hasiri-sugiru, sugiru describes the path of motion in the lexical compound in (2-a), but excessiveness in the syntactic compound in (2-b). sugiru is most commonly used in the second compositional sense (meaning “too much”), and it is only because of the locative basutei “bus stop” in (2-a) that we can identify it as a lexical compound. We identify the meaning of such JCVs using syntactic information gained from co-occurrence and verb complements.

2.2 Semantic classes

We classify the semantics of the V2 into three semantic classes: aspectual, spatial and adverbial (Niimi et al., 1987). We first describe the theoretical background behind this classification.

Our semantic classification is designed to reflect the semantic commonalities between compound verbs in Japanese and verb particle constructions in English. Our classification is thus congruous, e.g., with the findings of Lidner (1983) on the semantics of out and up in VPCs. Lidner (1983) identified two basic senses: a spatial meaning which describes the path of motion verbs, and an extended meaning which indicates a state of change in the form of an aspectual change or emotional change of state.

In this research, we divide Lindner’s “extended meaning” class into the aspectual and adverbial classes to enhance utility in MT and paraphrase applications, as illustrated by the following examples:

- **Spatial class**
  - MT: translate the V2 as a particle (keri-ageru ⇒ kick up)
  - Paraphrasing: paraphrase V1 using the te-form and a directional adverb
    (bōru o nage-ageta “throw up the ball” ⇒ bōru o nage-te ni ageta.)

- **Aspectual class**
  - MT: translate the V2 as a verb ((ame ga) huri-dasu ⇒ begin to rain)
  - Paraphrasing: paraphrase V2 using an auxiliary verb such as hazimeru “start”
    (sake o nomi-dasita “start to drink alcohol” ⇒ sake o nomi-hazimeta.)

- **Adverbial class**
  - MT: translate the V2 as an adverb (akire-kaeru ⇒ be thoroughly disgusted)
  - Paraphrasing: paraphrase the V2 with an adjective/adverb
    (yühaN o tabe-sugita “eat too much dinner” ⇒ yühaN o yūbuNni tabeta.)
We define a V2 to be ambiguous if it has uses which correspond to more than one of the three semantic classes. For example, dasu “evict” has two basic V2 meanings: a spatial meaning (e.g. tobi-dasu “jump out”, keri-dasu “kick out”), and an aspectual meaning (e.g. syaberi-dasu “start to talk” and tabe-dasu “start to eat”). Note that we exclude idiomatic JCVs like wari-dasu “count” from disambiguation, based on the reasoning that: (a) we have little chance of predicting their non-compositional semantics, and (b) they are non-productive and can thus be enumerated in a dictionary.

This research is based on 20 V2s with interpretational ambiguity, as shown in Table 1 with a representative example JCV for each semantic class the V2 belongs to.

<table>
<thead>
<tr>
<th>V2</th>
<th>Aspectual class</th>
<th>Spatial class</th>
<th>Adverbial class</th>
</tr>
</thead>
<tbody>
<tr>
<td>agaru</td>
<td>“go up”</td>
<td>yude-agaru “finish boiling”</td>
<td>tobi-agaru “jump up”</td>
</tr>
<tr>
<td>ageru</td>
<td>“lift”</td>
<td>yaki-ageru “finish baking”</td>
<td>hiki-ageru “pull up”</td>
</tr>
<tr>
<td>dasu</td>
<td>“put out”</td>
<td>take-dasu “start eating”</td>
<td>tobi-dasu “jump out”</td>
</tr>
<tr>
<td>iru</td>
<td>“enter”</td>
<td>osi-iru “break into”</td>
<td>hazi-iru “be ashamed”</td>
</tr>
<tr>
<td>kakaru</td>
<td>“hang”</td>
<td>ochi-kakaru “be dropping”</td>
<td>kiri-kakaru “slash at”</td>
</tr>
<tr>
<td>kakeru</td>
<td>“hang”</td>
<td>yomi-kakeru “start reading”</td>
<td>hanasi-kakeru “talk to”</td>
</tr>
<tr>
<td>kaeru</td>
<td>“go back”</td>
<td>huri-kaeru “look back”</td>
<td>osi-kaeru “be disgusted”</td>
</tr>
<tr>
<td>karesu</td>
<td>“send back”</td>
<td>uri-karesu “hit back”</td>
<td>yomi-karesu “read again”</td>
</tr>
<tr>
<td>kiru</td>
<td>“cuttrans”</td>
<td>tataki-kiru “chop off”</td>
<td>komari-kiru “be at a loss”</td>
</tr>
<tr>
<td>kireru</td>
<td>“cutintrans”</td>
<td>suri-kireru “wear out”</td>
<td>tabe-kireru “can eat all”</td>
</tr>
<tr>
<td>komu</td>
<td>“enter”</td>
<td>hari-komu “go into”</td>
<td>huke-komu “become old”</td>
</tr>
<tr>
<td>nuku</td>
<td>“pull out”</td>
<td>hiki-nuku “pull out”</td>
<td>osi-nuku “forget to say”</td>
</tr>
<tr>
<td>otosu</td>
<td>“drop”</td>
<td>kiri-otosu “cut off”</td>
<td>i-otosu “forget to say”</td>
</tr>
<tr>
<td>sugiru</td>
<td>“go past”</td>
<td>tōri-sugiru “go past”</td>
<td>tabe-sugiru “eat too much”</td>
</tr>
<tr>
<td>tateru</td>
<td>“standtrans”</td>
<td>ost-tateru “push up”</td>
<td>kaki-tateru “splash”</td>
</tr>
<tr>
<td>tutu</td>
<td>“standintrans”</td>
<td>orit-tatu “go down”</td>
<td>hurui-tatu “stop”</td>
</tr>
<tr>
<td>tobasu</td>
<td>“scatter”</td>
<td>tuki-tobasu “push aside”</td>
<td>sashimi-tobasu “bawl out”</td>
</tr>
<tr>
<td>tōsu</td>
<td>“pierce”</td>
<td>sas-tōsu “run through”</td>
<td>osi-tōsu “carry through”</td>
</tr>
<tr>
<td>takeru</td>
<td>“attach”</td>
<td>ost-takeru “press against”</td>
<td>sashimi-takeru “reprimand”</td>
</tr>
<tr>
<td>wataru</td>
<td>“cross”</td>
<td>hibiki-wataru “echo”</td>
<td>yuki-wataru “be widespread”</td>
</tr>
</tbody>
</table>

Table 1

Types of V2 ambiguity (trans = transitive, intrans = intransitive)

2.3 Semantically ambiguous V2s

We extracted data from the 1993 Mainichi Shinbun corpora (Mainichi Newspaper Co., 1993) in order to study actual patterns of occurrence/ambiguity of
Table 2
Token and type frequency of JCVs in the 1993 Mainichi corpus

<table>
<thead>
<tr>
<th>Verb type</th>
<th>Tokens</th>
<th>Types</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single word</td>
<td>1,349,419</td>
<td>4,355</td>
</tr>
<tr>
<td>Non-lexicalised JCV</td>
<td>106,409</td>
<td>7,819</td>
</tr>
</tbody>
</table>

JCVs. We tagged all the newspaper articles with the ChaSen splitter/tagger (Matsumoto et al., 2000), and extracted out all non-lexicalised JCVs (i.e. all JCVs which were split into their component verbs by ChaSen\(^2\)). From these, we excluded all JCVs found in the Shin Meikai dictionary (Kindaichi, 1999) so as to filter out idiomatic JCVs. In Table 2, we detail the number of single-word verbs (including lexicalised JCVs as listed in the ChaSen dictionary) and productive JCVs, in terms of both type and token instances in the 1993 Mainichi corpus.

Non-lexicalised (i.e. productive) JCVs account for only 7% of the total token count but 64% of all types. Of the 7,819 non-lexicalised JCV types, 4,730 (60% of all non-lexicalised tokens) were not contained in the Shin Meikai dictionary. This underlines the rich variety of JCV types and difficulty in processing JCVs by way of a pre-compiled dictionary. A total of 1,075 JCV types incorporating our 20 ambiguous V2s were contained in the non-lexicalised data not found in the dictionary.

3 Statistical approach

We use a statistical method as a first step in disambiguating JCVs. The statistical approach predicts the range of semantic classes that a given JCV can occur in by way of analysing patterns of combination of the V1 and V2 in other JCVs. This classification takes place at the JCV type level, such that in the case that the classifier predicts that a given JCV is monosemous (occurs in a unique semantic class), we are able to trivially tag all token instances of that JCV. In the case that the classifier predicts the JCV to be polysemous, we must rely on the rule-based method to perform context-sensitive sense disambiguation of token instances of that JCV, but can at least hope to constrain the sense inventory.

In order to build the JCV sense classifier, we first sense-tagged all token instances of the 1,075 JCVs found in the 1993 Mainichi corpus which were non-lexicalised, not in the Shin Meikai dictionary and also incorporated one

\(^2\) Note that ChaSen treats any JCV in its verb dictionary as a single word. This includes idiomatic JCVs such as tori-ageru “take away”, lexical JCVs such as kaki-ageru “write up” and syntactic JCVs such as yobi-tudukeru “keep calling”.

7
Table 3
Type-level breakdown across semantic types in the 1993 Mainichi data

<table>
<thead>
<tr>
<th>Type</th>
<th>Frequency</th>
<th>Proportion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aspects</td>
<td>563 (.520)</td>
<td></td>
</tr>
<tr>
<td>Spaces</td>
<td>265 (.245)</td>
<td></td>
</tr>
<tr>
<td>Adverbs</td>
<td>337 (.311)</td>
<td></td>
</tr>
</tbody>
</table>

Table 4
Matrix of JCV occurrence/semantic multiclass (A = aspectual, D = adverbial, S = spatial, DS = adverbial&spatial, ? = unobserved)

The 1,075 JCVs from the 1993 Mainichi corpus ranged over 551 V1s and 20 V2s, and occurred in 6 semantic multiclasss: aspect, space, adverb, aspect&space, space&adverb and aspect&adverb. For the purposes of classification, we represent the JCV data by way of a matrix describing the observed V1-V2 combinations, annotated with the semantic multiclass of each, as outlined in Table 3; in the case that a given V1-V2 combination is not observed, we tag it as ?. We extract a feature vector representation of each JCV by concatenating the column and row of the matrix in which it occurs, as outlined in (3) for the combination of V1_4 and V2_5 in Table 3. We hold out the class annotation of V1_i and V2_j by replacing the value for V1_i and V2_j by ?, as indicated by the boxes in (3).

(3) $V_{14} - V_{25}$ : $[A, S, ?, S, \ldots, DS, D, \ldots, ?, S, \ldots, ?]$
The classifier was learned using the TinySVM support vector machine learner. A polynomial kernel of order 2 was found to be the optimal configuration for the given task. We perform classification by way of three binary classifiers, one for each atomic semantic class (i.e. aspectual, spatial and adverbial). In independent research, we tested various classifier configurations with the TiMBL memory-based learner (Daelemans et al., 2003) and found the binary classifier suite architecture to be optimal. TinySVM was found to marginally outperform TiMBL, and we thus report only the results for TinySVM in this paper.

We tested two basic feature vector formats: semantic categorisation and collocational categorisation. Semantic categorisation is a slight variant on the format presented in (3) above, with the semantic class of each observed JCV represented by way of a binary-valued triple representing the breakdown of the semantic multiclass into its component semantic classes (e.g. $A = [1, 0, 0]$, $DS = [0, 1, 1]$ and $? = [?, ?, ?]$). In collocational categorisation, on the other hand, we simply record the observed collocational pattern of the V1 and V2 making up the given JCV, replacing each semantic class in (3) with a 1 (for observed), and each ? with a 0 (for unobserved) except for values $V1_i$ and $V2_j$ which we leave as ? to inform the classifier of the identity of the $V1_i$ and $V2_j$ in the JCV we are attempting to classify. The collocational feature vector for $V1_4 - V2_5$ is thus as follows:

\[
V1_4 - V2_5 : [1, 1, 0, ?, 1, \cdots, 0, 1, \cdots, 0, ?, 1, \cdots, 0]
\]

Collocational categorisation has the obvious advantage over semantic categorisation that it is trivially scalable to novel V1s and V2s, as the method does not rely on semantic annotation of newly observed JCVs. In the case of semantic categorisation, on the other hand, any addition of a new row or column to the JCV matrix requires extra annotation of observed JCVs occurring in that same row or column. In this sense, collocational categorisation of the JCV data is the more versatile of the two proposed methods, and better suited to JCV productivity.

Note that it is possible to hybridise the two categorisation methods, e.g. in applying collocational categorisation to the V1 (column) data and semantic categorisation to the V2 (row) data, an avenue we explore below.

\[^3\text{http://cl.aist-nara.ac.jp/~taku-ku/software/TinySVM/index.html}\]
Table 5
Type-level evaluation of the statistical method ($A = \text{accuracy}$, $F = \text{F-score}$)

<table>
<thead>
<tr>
<th>Feature categorisation</th>
<th>Aspect $A$</th>
<th>Aspect $F$</th>
<th>Spatial $A$</th>
<th>Spatial $F$</th>
<th>Adverb $A$</th>
<th>Adverb $F$</th>
<th>Total $A$</th>
<th>Total $F$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coll × Coll</td>
<td>0.933</td>
<td>0.897</td>
<td>0.851</td>
<td>0.680</td>
<td>0.914</td>
<td>0.917</td>
<td>0.899</td>
<td>0.831</td>
</tr>
<tr>
<td>Coll × Sem</td>
<td><strong>0.935</strong></td>
<td><strong>0.900</strong></td>
<td>0.853</td>
<td>0.689</td>
<td>0.922</td>
<td>0.925</td>
<td>0.903</td>
<td><strong>0.838</strong></td>
</tr>
<tr>
<td>Sem × Coll</td>
<td>0.932</td>
<td>0.894</td>
<td>0.852</td>
<td>0.673</td>
<td>0.924</td>
<td>0.926</td>
<td>0.903</td>
<td>0.831</td>
</tr>
<tr>
<td>Sem × Sem</td>
<td>0.924</td>
<td>0.881</td>
<td><strong>0.860</strong></td>
<td><strong>0.697</strong></td>
<td><strong>0.928</strong></td>
<td><strong>0.929</strong></td>
<td><strong>0.904</strong></td>
<td><strong>0.836</strong></td>
</tr>
<tr>
<td>Baseline</td>
<td>0.520</td>
<td>—</td>
<td>0.755</td>
<td>—</td>
<td>0.689</td>
<td>—</td>
<td>0.480</td>
<td>—</td>
</tr>
</tbody>
</table>

3.1 Evaluation

We evaluated the statistical method by way of 10-fold stratified cross validation, using the semantic ($\text{Sem}$) and collocational ($\text{Coll}$) categorisation methods for each of the V1 and V2 data, producing the four configurations given in Table 3.1; $\text{Coll} \times \text{Sem}$, e.g., indicates that we employ collocational categorisation for the V1 data and semantic categorisation for the V2 data. We evaluate each combination of feature categorisation individually across the three semantic classes and also cumulatively, according to classification accuracy ($A$) and F-score ($F$). The highest accuracy and F-score each data category is presented in **bold** in Table 3.1. In each case, we also present a baseline accuracy, based on a majority-class strategy (e.g. in the case of the spatial class, this corresponds to negative classification, and in the case of combined classification, this corresponds to the aspect multiclass).

The total accuracy is around 90% for each classifier configuration—well above the baseline accuracy of 48%—while the total F-score is around 83% in each case; none of the differences are statistically significant. Given that JCV productivity is in terms of the scope for a given V2 to take different V1s, we can make the case that $\text{Coll} \times \text{Sem}$ is the optimal classifier configuration, as it has the highest overall F-score and requires only collocation information in order to expand the JCV matrix to include novel V1s. Alternatively, if we are interested in fully open-class JCV disambiguation beyond the bounds of the 20 V2s targeted in this research, $\text{Coll} \times \text{Coll}$ should be the model of choice as it requires only information on patterns of co-occurrence and no additional annotation of JCV data according to semantic class.

Looking to the breakdown in classifier performance across the three semantic classes, we see no significant variation in classifier performance for a given semantic class, but considerable variation across the different classes. The spatial semantic class appears to be the hardest to predict, with the diminished F-score in each case occurring due to slightly diminished precision and a dra-
matic drop in recall. The reason for this is the difficulty in determining spatial semantics without context, e.g. *umi e kogi-dasu* “row far out to sea” has spatial semantics, while *totuzen kogi-dasu* “begin to row suddenly” has aspectual semantics.

*Collx Coll* predicts that 101 (6.0%) of the JCV tokens are ambiguous and thus require token-level disambiguation. For the remaining 974 examples (94.0%), it is possible to sense-tag all token instances using the unique class returned by the classifier.

4 Rule based approach

In the rule-based approach, we construct disambiguation rules in a two-step process: (1) identify the meaning of the V1, and (2) classify the JCV and cluster according to the semantics of the V1 and its verb complements. We manually construct V2 sense disambiguation rules based on the obtained semantic and syntactic information.

The primary means of disambiguating JCV sense is the semantics of the V1. We base our semantic classification on Ruigo Shin Jiten (Oono and Hamanishi, 1989): e.g., *musu* “steam” is classified as *suizyi* “kitchen work” and *nageru* “throw” as *dageki* “throw and hit”; alternatively, a verb may be classified according to multiple categories, e.g. *utu* “hit” is classified as both *dageki* “throw and hit” and *suizyi* “kitchen work”. These features can be used to distinguish the aspectual use of *ageru* in *musi-ageru* “finish steaming” from the spatial use in *nage-ageru* “throw into the air”. Ruigo Shin Jiten is organised into three levels and constitutes 1000 categories. The labels at the second level, which include 60 categories for verbs, are used in assigning semantics to the V1. If we have difficulty in discriminating the V1 semantics using this level of sense granularity, we use the labels from the third level.

4.1 Information for Disambiguation Rules

We analysed the semantics of all 551 V1s occurring in the 1,075 JCVs extracted out of the 1993 Mainichi corpus. We complemented the V1 semantics with contextual information, namely valence information and the semantic class of noun complements of the JCV, based on the IPAL verb dictionary (IPA, 1987). The IPAL verb dictionary defines the meaning of verbs using valence patterns and assigns a semantic feature from Ruigo Shin Jiten to each entry. Contextual disambiguation takes place in two steps: (1) disambiguate the meaning of the V1 according to the IPAL verb dictionary; and (2) identify semantic and
syntactic information for use in the disambiguation rules.

First, we extract the noun complements of each JCV token, and identify the corresponding V1 valency pattern in the IPAL dictionary. For example, in the case of uti-ageru “finish making” in a context like kare-wa udon-o uti-ageta “He finished making buckwheat noodles”, we try to find the valence entry of utu “hit” which is compatible with the phrasal complements kare-wa “he-TOP” and udon-o “buckwheat noodles-ACC”. Once we have identified the valence entry for utu “hit” which corresponds to this usage, a label like seisun “production” is selected as the semantic classification for utu “hit”. The second step is then to classify the JCV into one of the three semantic classes based on the criteria as defined in Section 2.2, and to cluster JCVs of the same semantic type together according to the semantic and syntactic features of the V1. For example, JCVs such as yude-ageru “finish boiling”, musi-ageru “finish steaming” and yaki-ageru “finish baking” are all classified into the aspectual class, and all have the common V1 semantic feature suizyi “kitchen work”. Noun complements of the V1 and their semantic features can also be used in disambiguating the meaning of the V2. For instance, unazuku “nod” has the single meaning of “agreement”. The combination of unazuku “nod” and kakeru “hang” (i.e. unazuki-kakeru), however, is ambiguous between an aspectual meaning (e.g. kare-wa sono kotoba-ni unazuki-kaketa “he was about to nod at what was being said”) and a spatial meaning (e.g. kare-ni unazuki-kaketa “I nodded at him”). In this case, we combine the V1 semantics with syntactic information and the semantics of the noun complements in disambiguating the JCV sense.

4.2 Disambiguation Rules

In order to construct the disambiguation rules, the JCVs were classified into two groups using the results of the analysis from Section 4.1. The rules in the first group (semantic rules) are based purely on the V1 semantics. We built disambiguation rules based on the V1 semantic classification of the Ruigo Shin Jiten. The rules are composed of the V1 semantic class, the lexical form of the V2 and the corresponding semantic class of the V2.

The second rule group (syntactico-semantic rules) involves V1 semantics and also syntactic and semantic information on verb complements. Examples of the different disambiguation rule types are as follows:

• Semantic rules
  - Rule 1: IF V1 IS A cooking verb and V2 is ageru THEN V2 is aspectual
    E.g. yude-ageru “boil-raise” = yuderu-koto-o oeru “finish boiling”
  
• Rule 2: IF V1 IS A operation verb and V2 is ageru THEN V2 is spatial
E.g. *uti-ageru* “hit-raise” = *utte-ageru* “hit upwards”

- **Rule 3:** IF V1 IS A emotion verb and V2 is *agaru* THEN V2 is **adverbial**
  E.g. *hurue-agaru* “tremble-go up” = *hizyouni hurueru* “tremble violently”

• **Syntactico-semantic rule**
  - **Rule 4:** IF V1 IS A action verb and N1 (V1’s subject) IS A human and N2 (V2’s dative) IS A human and V2 is *kakeru* THEN V2 is **aspectual**
  E.g. *kare-wa kanozyo-ni unazuki-kaketa* “he-TOP she-DAT nod-hang” = *kare-wa kanozyo-ni mukatte unazuita* “He nodded at her”

We constructed rules based on the 1,075 JCVs extracted from the 1993 Mainichi corpus. The final rule set consists of 113 semantic rules and 34 syntactico-semantic rules.

It is important to realise that, in the current formulation, the rules are both formulated and applied manually by a human oracle. That is, all rules rely on manual disambiguation of V1 sense, and the syntactico-semantic rules additionally require manual determination of phrasal structure and disambiguation of the sense of each complement. Clearly automating the processes of both rule construction and application would be possible, but inevitably lead to errors. For our present purposes, we use the rule-based method in determining an upper bound on disambiguation performance, to benchmark the statistical method against.

### 4.3 Evaluation

In order to evaluate the rule-based method and the token-level accuracy of the statistical method, we extracted up to 5 token occurrences of each of the 1,075 JCVs types from the 1994 Mainichi corpus (Mainichi Newspaper Co., 1994), once again based on the output of ChaSen. Recall that the 1993 Mainichi corpus was used to extract out the original JCVs and for rule development, such that these token instances constitute held-out test data. A total of 622 of our JCV types were observed in the 1994 Mainichi corpus, at a combined token count of 1,515 (406 **aspectual**, 420 **spatial** and 689 **adverbial**).

We present an analysis of token-level classification accuracy over the test data in Table 6 for the pure rule-based system (without using the statistical method as a sense filter), a baseline method (corresponding to tagging all token instances as the majority class of **adverbial**), and also the various configurations of the statistical method. In the case of the statistical method, we present the accuracy for monosemous JCVs (i.e. JCVs which are classified according to a unique semantic class) and also polysemous JCVs, with token-level disambiguation performed either randomly or according to the rule-based method.
<table>
<thead>
<tr>
<th></th>
<th>Monoseous</th>
<th>Polyseous</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pure rule-based</td>
<td>—</td>
<td>—</td>
<td>.960</td>
</tr>
<tr>
<td>Coll×Coll</td>
<td>.853</td>
<td>.500/.949</td>
<td>.825/.860</td>
</tr>
<tr>
<td>Coll×Sem</td>
<td><strong>.885</strong></td>
<td>.500/.945</td>
<td>.857/.<strong>889</strong></td>
</tr>
<tr>
<td>Sem×Coll</td>
<td>.879</td>
<td>.500/.982</td>
<td>.865/.883</td>
</tr>
<tr>
<td>Sem×Sem</td>
<td>.862</td>
<td>.500/ <strong>.983</strong></td>
<td>.831/.867</td>
</tr>
<tr>
<td>Baseline</td>
<td>—</td>
<td>—</td>
<td>.455</td>
</tr>
</tbody>
</table>

Table 6
Token-level accuracy over the 1994 Mainichi corpus

Similarly for the overall accuracy, we present figures based on first random then rule-based disambiguation.

The pure rule-based method significantly outperforms the various configurations of the statistical method, all of which are vastly superior to the baseline. The high level of monosemous outputs from the statistical method means that there is relatively little difference between random and rule-based disambiguation, suggesting that the high manual overhead it entails is perhaps not commensurate with its impact on performance. The best-performing of the statistical methods is Coll×Sem, mirroring the type-level results in Table 3.1.

The errors produced by the pure rule-based method can be divided into three types: gaps in rules, inappropriate assignment of semantic features and exceptions.

Gaps in rules occurred when a particular combination of V1 semantics and V2 (and optionally syntactic and noun semantic information) were not found in our rules due to that interpretation type not being observed in the 1993 Mainichi data.

An example of inappropriate assignment of V1 semantic features is oru “fold”, which is classified as henkei “transformation” but also denotes a verb of creation. The difference between the classes henkei “transformation” and seisan “production” is important as ori-agaru “complete folding” should be in the aspectual class, but is misclassified as spatial due to the existence of JCVs with “transformation” type V1 semantics such as maki-agaru “be rolled up”. We consider this a limitation of the semantic resources used in this research.

The final type of error was exceptions, where an unusual V1 usage leads to misclassification. For example, tukuru “make” is assigned the seisan “production” class, such that tukuri-komu “make” is misclassified as spatial due to the existence of JCVs such as ami-komu “knit” and nui-komu “sew”. The correct
semantic classification of *tukuri-komu* “make”, however, is adverbial, which must be registered in the dictionary as an exception.

5 Discussion and Future Work

We have proposed two methods for disambiguating JCVs: a statistical and rule-based method. The statistical method makes use of collocational or semantic information about different V1-V2 combinations in classifying JCVs at the type level. The rule-based method, on the other hand, relies on manual semantic analysis of the V1 and also verb complements in classifying JCVs at the token level. We presented a two-step system where the output of the statistical method is fed into the rule-based method, and found that the pure rule-based method significantly outperformed the combined method at 96% accuracy. At the same time, we established that the statistical method generally produces a monosemous type-level classification, such that token-level sense tagging is a trivial task, and that even if we randomly select a sense in the case of polysemy, the token-level accuracy is a creditable 86%. It would thus appear that semantics can offer improvements in token-level accuracy, but that at the same time, the statistical method can be used successfully for fully-automated token-level sense-tagging.

In future work, we are first and foremost interested in automating the rule-based method to determine whether it will be possible to boost the performance of the statistical method through automatic means, or indeed whether a standalone token-level classification method of the nature of the rule-based method will be superior in performance. In the current setup, the rule-based method involves word sense disambiguation of both the V1 and each verb complement, as well as dependency parsing of the input strings. Given the performance limitations of present-day word sense disambiguation techniques and also the high productivity of JCVs, it is doubtful that we could expect to surpass our current statistical method, but this remains an item for further investigation.

There are a number of additional syntactic features that we could attempt to incorporate into the classification process. The valence of the V1 and V2 could be used in our type-level JCV classification, as mismatches in valence tend to lead to semantic anomalies. While this would lead to complications in terms of scalability (i.e. we would run the risk of not knowing the valence of novel V1s and V2s), it appears an avenue worth pursuing in future research. Similarly, there appear to be subtle interactions between the lexical/syntactic categorisation of JCVs and their semantics which we could make use of (Hashimoto, 2003).
We are also very interested in applying the methods proposed here to the task of sense classification of English verb particles, and also in the MT of JCVs.

References


Sag, I. A., T. Baldwin, F. Bond, A. Copestake, and D. Flickinger (2002). Multiword expressions: A pain in the neck for NLP. In Proc. of the 3rd Interna-
