Disambiguating Japanese Compound Verbs

Kiyoko Uchiyama [a,b], Timothy Baldwin [c,b], Shun Ishizaki [a]

[a] *Graduate School of Media and Governance*
Keio University
5322 Endo, Fujisawa-shi, Kanagawa 252-8520 Japan

[b] *Center for the Study of Language and Information*
Stanford University
210 Panama Street, Stanford, CA 94305 USA

[c] *Department of Computer Science and Software Engineering*
University of Melbourne, Victoria 3010 Australia

Abstract

The purpose of this study is to disambiguate Japanese compound verbs (JCVs) using two methods: (1) a statistical sense discrimination method based on verb-combinatoric information, which feeds into a first-sense statistical sense disambiguation method, and (2) a manual rule-based sense disambiguation method which draws on argument structure and verb semantics. In evaluation, we found that the rule-based method outperformed the statistical method at 94.6% 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 82.6%, without making use of syntactic or lexical semantic knowledge.

Key words: word sense disambiguation, word sense discrimination, 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,

Email addresses: `kiyoko@sfc.keio.ac.jp` (Kiyoko Uchiyama), `tim@csse.unimelb.edu.au` (Timothy Baldwin), `ishizaki@sfc.keio.ac.jp` (Shun Ishizaki).

Preprint submitted to Elsevier Science *6 January 2005*

Uchiyama, Kiyoko, Timothy Baldwin and Shun Ishizaki (2005) Disambiguating Japanese Compound Verbs, Computer Speech and Language, Special Issue on Multiword Expressions, Volume 19, Issue 4, pp. 497-512.

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 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 in *osi-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 *teNtō-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. *(bōru o) nage-ageru* “throw (a ball) up”
(bōru o) keri-ageru “kick (a ball) up”
b. *(yasai o) yude-ageru* “finish boiling (vegetables)”
(yasai o) musi-ageru “finish steaming (vegetables)”
c. *donari-ageru* “shout”
odosi-ageru “threaten”

Ageru “raise” has multiple meanings when it appears in the V2 position of a JCV. In (1-a), spatial compound verbs are formed in combination with a V1 verb of motion. Conversely, in (1-b), aspectual compound verbs are formed in combination with a V1 cooking verb. In addition, in (1-c), adverbial compound verbs are formed in combination with an emotional-type V1. See Section 2.2 below for detailed definitions of the semantic classes.

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 are interpreted differently according to the semantic properties of the V1.

JCVs have received a moderate amount of attention in the fields of linguistics and natural language processing. In linguistics, JCVs have been studied mainly

[1] The following abbreviations are used in glosses in this paper: TOP = topic, SUBJ subject, ACC = accusative and DAT = dative.

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 the V1 and V2, we claim that it is necessary to analyse the JCV compositionally based on the individual semantics of the V1 and V2.

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.

Uchiyama and Ishizaki (2003) proposed a manual rule-based approach to type-level semantic class labelling which provided the foundation for this research. We reuse Uchiyama and Ishizaki's three-way semantic classification of JCVs (see Section 2.2), and the same lexical resources in our rule set (namely IPAL and Ruigo Shin Jiten — see Section 4). Our rule set differs in that it is intended for token-level semantic labelling, and covers 20 rather than the original 10 ambiguous V2s. Additionally, we carry out more detailed evaluation of the rule set, at the token rather than type level to estimate the performance of the method over a document or other text stream.

We propose a disambiguation method which identifies the meaning of JCVs using two basic steps: (1) a statistical approach which determines the type-level sense inventory of a given V1-V2 combination independent of context, and applies a naive first-sense strategy in disambiguating token-level instances of polysemous JCVs; and (2) a pure 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 first perform word sense discrimination, i.e. identify the range of meanings a given JCV type can take. The bulk 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, and by adopting a simple first-sense strategy conditioned on the V2 it is possible to perform rough-and-ready word sense disambiguation. 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 learned from the collocational data.

The rule-based approach 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.

This research has applications in language understanding tasks and MT, notably 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)). We claim that VPCs 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 aspectual (*write up a paper*, 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. Section 3 describes the statistical JCV sense disambiguation method. Section 4 details the pure rule-based for disambiguating JCV sense at the token level. Section 5 then evaluates the two methods.

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* "finish boiling") 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.

Martin (1975) and Tsujimura (1996) explain the difference between syntactic JCVs and lexical compounds based on the concepts of grammatical properties, meaning and formation.

- **Syntactic compounds**
  - · The argument structure and syntactic distribution of arguments and adjuncts of the V1 is preserved in the JCV;
  - · The meaning of the compound is wholly predictable from the component verbs;
  - · JCV formation is productive or semi-productive.
- **Lexical compounds**
  - · Lexical compounds do not always preserve the grammatical properties of the V1;
  - · The meaning of the compound cannot always be predicted from the components;
  - · JCV formation is restricted and combinatory possibilities are pre-determined with fixed meanings.

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.

In this research, we focus on JCVs which contain an ambiguous V2 (as in (1) above). Semantic constraints govern the V1-V2 compounding process, such that while *yude-ageru* "finish boiling" 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 <u>drove past</u> the bus stop.

b. *Kare wa siai tyū hasiri-sugita.*
He <u>ran too much</u> during the game.

In *hasiri-sugiru* above (past tense = *hasiri-sugita*), *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 VPCs in English. Our classification is thus congruous, e.g., with the findings of Lindner (1983) on the semantics of *out* and *up* in VPCs. Lindner 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:

- **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.*)
- **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" ⇒ *bouru o nage-te ue ni ageta.*)
- **Adverbial class**
  - · MT: translate the V2 as an adverb (*akire-kaeru* ⇒ *be thoroughly disgusted*)
  - · Paraphrasing: paraphrase the V2 with an adjective/adverb (*yūhaN o tabe-sugita* "eat too much dinner" ⇒ *yūhaN o zyūbuNni tabeta.*)

2.3 Semantically ambiguous V2s

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

V2	***Aspectual class***	***Spatial class***	***Adverbial class***
agaru "go up"	*yude-agaru* "finish boiling"	*tobi-agaru* "jump up"	*hurue-agaru* "be terrified"
ageru "raise"	*yaki-ageru* "finish baking"	*hiki-ageru* "pull up"	
dasu "put out"	*tabe-dasu* "start eating"	*tobi-dasu* "jump out"	
iru "enter"		*osi-iru* "break into"	*hazi-iru* "be ashamed"
kakeru "hang$_{\text{trans}}$"	*yomi-kakeru* "start reading"	*hanasi-kakeru* "talk to"	
kakaru "hang$_{\text{intrans}}$"	*ochi-kakaru* "be dropping"	*kiri-kakaru* "slash at"	
kaeru "go back"		*huri-kaeru* "look back"	*akire-kaeru* "be disgusted"
kaesu "send back"		*uti-kaesu* "hit back"	*yomi-kaesu* "read again"
kiru "cut$_{\text{trans}}$"		*tataki-kiru* "chop off"	*komari-kiru* "be at a loss"
kireru "cut$_{\text{intrans}}$"		*suri-kireru* "wear out"	*tabe-kireru* "can eat all"
komu "enter"		*hairi-komu* "go into"	*huke-komu* "become old"
nuku "pull out"		*hiki-nuku* "pull out"	
otosu "drop"		*kiri-otosu* "cut off"	*ii-otosu* "forget to say"
sugiru "go past"		*tōri-sugiru* "go past"	*tabe-sugiru* "eat too much"
tateru "stand$_{\text{trans}}$"		*osi-tateru* "push up"	*kaki-tateru* "splash"
tatu "stand$_{\text{intrans}}$"		*ori-tatu* "go down"	*hurui-tatu* "stir"
tobasu "scatter"		*tuki-tobasu* "push aside"	*sikari-tobasu* "bawl out"
tōsu "pierce"		*sasi-tōsu* "run through"	*osi-tōsu* "carry through"
tukeru "attach"	*tabe-tukeru* "eat habitually"	*osi-tukeru* "press against"	*sikari-tukeru* "reprimand"
wataru "cross"		*hibiki-wataru* "echo"	*yuki-wataru* "be widespread"

Table 1
Types of V2 ambiguity (trans = transitive, intrans = intransitive)

in Table 1 with a representative example JCV for each semantic class the V2 belongs to.

2.4 Extraction of JCVs

We extracted data from the 1993 Mainichi Shinbun corpus (Mai1993 hereafter: Mainichi Newspaper Co. (1993)) in order to study actual patterns of occurrence/ambiguity of JCVs. We tagged Mai1993 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]). In Table 2, we detail the number of single-word verbs (i.e. JCVs lexicalised in the ChaSen dictionary) and productive JCVs, in terms of both type and token instances in Mai1993.

Non-lexicalised (i.e. productive) JCVs account for only 7% of the total token count but 64% of all types. From these, we excluded all JCVs found in the

[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".

Verb type	***Tokens***	***Types***
Single word	1,349,419	4,355
Non-lexicalised JCV	106,409	7,819

Table 2
Token and type frequency of JCVs in MAI1993

Shin Meikai dictionary (Kindaichi, 1999) so as to filter out idiomatic JCVs.[3] This resulted in 4,730 types (60% of all non-lexicalised tokens), underlining 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 this final dataset.

3 Statistical sense disambiguation

We perform statistical sense disambiguation of JCVs through a 2-step process. First, we perform **word sense discrimination** in the sense of Schütze (1998) to identify the range of senses that a given JCV type can take. E.g., for *uti-ageru*, we would hope to identify the fact that it can occur with either spatial or aspectual semantics. Second, we perform **word sense disambiguation** in tagging each token occurrence of that JCV according to a first-sense classifier conditioned on the V2 and constrained by the sense inventory identified in the first step. E.g., the predominant sense for *ageru* may be identified as spatial, such that token instances *udoN-o uti-ageru* "finish making wheat noodles" [aspectual] and *roketto-o uti-ageru* "launch a rocket" [spatial] would both be disambiguated according to the spatial class.

The primary motivation for not tackling JCV sense disambiguation directly, i.e. not building a token-level semantic tagger, is that the majority of JCVs are monosemous (i.e. are classified according to a unique semantic class), obviating the need for explicit word sense disambiguation. Additionally, this leads to a very robust disambiguation method which is able to cope with novel V1-V2 combinations, immune to the effects of varying domain/context and requires no prior lexico-syntactic knowledge of the V1.

3.1 Sense discrimination

Word sense discrimination is performed by way of analysing patterns of combination of the V1 and V2 in other JCVs. In order to build the JCV sense

[3] Note that for these idiomatic JCVs, we are still able to arrive at an interpretation based on the dictionary definition.

	Aspectual	***Spatial***	***Adverbial***
Type frequency	181 (.168)	333 (.310)	561 (.522)

Table 3
Type-level breakdown across semantic types in Mai1993

		V2								
		$V2_1$	$V2_2$	$V2_3$	$V2_4$	$V2_5$	$V2_6$	$V2_7$	$\dots$	$V2_j$
	$V1_1$					A				
	$V1_2$					S				
	$V1_3$					?				
V1	$V1_4$	D	...		?	A	S	...		?
	$V1_5$					S				
	$\vdots$					$\vdots$				
	$V1_i$					DS				

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

discrimination classifier, we first sense-tagged all token instances of the 1,075 JCVs found in Mai1993 which were not in the ChaSen or Shin Meikai dictionaries and also incorporated one of the 20 V2s targeted in this research. From these, we derived a type-level sense classification of each JCV type according to the Cartesian product of the three semantic classes (aspectual, spatial and adverbial). Of the 1,075 JCVs, a modest 83 were found to be polysemous, an example of which is *uti-ageru* which occurs with both spatial and aspectual semantics, as outlined above. In the remainder of cases, however, simple specification of the V1 was all that was required to disambiguate the V2 sense. We present the number of JCV types corresponding to each semantic class in Table 3, along with the overall proportion of JCVs this represents.

The 1,075 JCVs from Mai1993 ranged over 551 V1s and 20 V2s, and occurred in 6 semantic multiclasses: aspect, space, adverb, aspect&space, space&adverb and aspect&adverb.[4] 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 4; 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 indicated in (3) for the combination of $V1_4$ and $V2_5$ in Table 4. 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).

[4] It is interesting to note that no JCV was found which occurred with all three senses, and we are agnostic as to whether such JCVs exist.

(3) $$V1_4 - V2_5 :\ [\overbrace{\mathsf{D},\cdots,?,\boxed{?},\mathsf{S},\cdots,?}^{V2},\overbrace{\mathsf{A},\mathsf{S},?,\boxed{?},\mathsf{S},\cdots,\mathsf{DS}}^{V1}]$$

The sense discrimination classifier was learned using the TinySVM support vector machine learner.[5] 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 separate research (Uchiyama and Baldwin, 2004), we tested a wide range of classifier configurations with the TiMBL memory-based learner (Daelemans et al., 2003), but report only the results for TinySVM in this paper as it was found to marginally outperform TiMBL.

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. $\mathsf{A} = [1, 0, 0]$, $\mathsf{DS} = [0, 1, 1]$ 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:

(4) $$V1_4 - V2_5 :\ [\overbrace{1,\cdots,0,\boxed{?},1,\cdots,0}^{V2},\overbrace{1,1,0,\boxed{?},1,\cdots,1}^{V1}]$$

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.

[5] `http://cl.aist-nara.ac.jp/~taku-ku/software/TinySVM/index.html`

Feature	***Aspect***		***Spatial***		***Adverb***		***Total***	
categorisation	***A***	***F***	***A***	***F***	***A***	***F***	***A***	***F***
Coll×Coll	.933	.897	.851	.680	.914	.917	.899	.831
Coll×Sem	**.935**	**.900**	.853	.689	.922	.925	.903	**.838**
Sem×Coll	.932	.894	.852	.673	.924	.926	.903	.831
Sem×Sem	.924	.881	**.860**	**.697**	**.928**	**.929**	**.904**	.836
Baseline	.832	—	.690	—	.522	—	.681	—

Table 5
Evaluation of the statistical word sense discrimination method over MAI1993 (***A*** = accuracy, ***F*** = F-score)

3.1.1 Evaluation

We evaluated the sense discrimination method by way of 10-fold stratified cross validation over the MAI1993 data, using the semantic (*Sem*) and collocational (*Coll*) categorisation methods for each of the V1 and V2 data, producing the four configurations given in Table 6; *Coll*×*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[6] (F). The highest accuracy and F-score of each data category is presented in **bold** in Table 6. 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 adverb multiclass).

The total accuracy is around 90% for each classifier configuration—well above the baseline accuracy of 68%—while the total F-score is around 83% in each case; none of the differences are statistically significant. Based on this, we lose nothing in terms of performance by basing the classification solely on collocation information (i.e. using the *Coll*×*Coll* classifier), but gain considerably in terms of versatility and scalability.

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 dramatic 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 spa-

[6] F-score was calculated according to the formula $\frac{2PR}{R+P}$, where P is precision and R is recall.

tial semantics, while *totuzen kogi-dasu* "begin to row suddenly" has aspectual semantics.

Coll×*Coll* predicts that 101 (6.0%) of the JCV types are ambiguous and thus require token-level sense 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.

3.2 First-sense word disambiguation

Having identified the range of semantic classes a given JCV type occurs in, we next look to token-level sense disambiguation. As observed above, 94.0% of JCVs are classified as monosemous, and for these JCVs, we can simply tag all token instances according to the unique semantic class. For the polysemous JCVs (the remaining 6.0% of the data), we employ a simple first-sense strategy to disambiguate token instances thereof. We identified the first sense for each V2 by manually identifying the distribution of senses over a random sample of 100 token instances involving that V2 in MAI1993. Based on this, we generate an individual sense ranking for each V2, and disambiguate all token instances of a given JCV based on the highest-ranking V2 sense which is found within the subset of senses identified by the sense discrimination classifier. As an example of this process, assume the sense ranking for the V2 *ageru* is (aspectual,spatial,adverbial)—i.e., aspectual is the predominant sense, followed by spatial and adverbial—and the classification for *hineri-ageru* "twist up" is spatial+adverbial. The first sense for *ageru* is spatial, therefore, we tag all token instances of *hineri-ageru* as spatial.

The first-sense strategy is a well-established baseline in word sense disambiguation research which has been shown to be surprisingly effective over a variety of datasets (Kilgarriff, 2004; McCarthy et al., 2004). As a baseline, it can undoubtedly be improved upon, and the rule-based sense disambiguation method (Section 4) gives us some insight into types of features we could employ in such an endeavour. Given the relative infrequency of polysemous JCVs, however, the first-sense disambiguation method seems adequate for our immediate needs (see Section 5 for further discussion).

4 Rule-based sense disambiguation

In the rule-based approach, we construct disambiguation rules in a two-step process, in the manner of Uchiyama and Ishizaki (2003): (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 Features employed in the disambiguation rules

We analysed the semantics of all 551 V1s occurring in the 1,075 JCVs extracted out of MAI1993. 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 wheat 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* "wheat noodles-ACC". Once we have identified the valence entry for *utu* "hit" which corresponds to this usage, a label like *seisaN* "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 *saNsei* "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 cooking_verb(V1) AND V2 = *ageru* THEN aspectual(V2)
    E.g. *yude-ageru* "boil-raise" = *yuderu-koto-o oeru* "finish boiling"

  - · Rule 2: IF operation_verb(V1) AND V2 = *ageru* THEN spatial(V2)
    E.g. *uti-ageru* "hit-raise" = *utte-ageru* "hit upwards"

  - · Rule 3: IF emotion_verb(V1) AND V2 = *agaru* THEN adverbial(V2)
    E.g. *hurue-agaru* "tremble-go up" = *hizyouni hurueru* "tremble violently"

- **Syntactico-semantic rule**
  - · Rule 4: IF action_verb(V1) AND human(N1 [= SUBJ]])) AND human(N2 [= DAT]) AND V2 = *kakeru* THEN spatial(V2)
    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 MAI1993. The final rule set consists of 113 semantic rules and 34 syntactico-semantic rules. This is an extension of the rule set constructed by Uchiyama and Ishizaki (2003), also based on MAI1993.

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.

5 Evaluation

In order to evaluate the statistical and rule-based disambiguation methods, we employed Mai1994$_{\text{RAND}}$, a random selection of 500 JCV-containing sentences taken from the 1994 Mainichi corpus (Mainichi Newspaper Co., 1994). The total number of JCV types represented in Mai1994$_{\text{RAND}}$ was 304, with a median JCV token frequency of 1 and a maximum JCV token frequency of 5. Of the 304 JCV types, 198 also occurred in Mai1993 and 106 were novel occurrences. As the sense rankings used in the statistical disambiguation method and the rules used in the rule-based disambiguation method are both based on Mai1993, this dataset represents a held-out test set for evaluation purposes. This dataset thus provides an estimate of the accuracy of the two methods over JCVs occurring in open text.

For the statistical disambiguation method, we first took the union of the collocational statistics in Mai1993 and Mai1994 and the manual sense classifications from Mai1993, and learned a sense discrimination classifier based on the Coll×Coll method. We evaluated the output of this classifier over the 304 JCV types according to three categories: (1) those JCVs seen in both Mai1993 and Mai1994 (**Seen**); (2) those JCVs novel to Mai1994 (**Unseen**); and (3) all 304 JCVs (**Total**). The results are presented in Table 6.

Table 6 reveals a number of interesting tendencies. First, performance over seen JCVs is consistently better than over unseen JCVs. This is not surprising when one considers that seen JCVs exist as labelled exemplars in the training data, where unseen JCVs do not. This then begs the question why we did not achieve 100% classification accuracy for the seen JCVs, the reason for which is that there are significant divergences in classifications across the two datasets, symptomatic of differences in JCV usage rather than a lack of linear separability in the feature space. Compared to the results for cross-validation over Mai1993 (see Table 6), the performance for seen JCVs is high overall (0.923 vs. 0.831 F-score). The performance for unseen JCVs, on the other hand, is slightly lower (0.755 vs. 0.831 F-score), largely because unseen JCVs tended to be formed from novel V1s, for which collocational data is limited.

JCV	***Aspect***		***Spatial***		***Adverb***		***Total***	
category	***A***	***F***	***A***	***F***	***A***	***F***	***A***	***F***
Seen	.939	.887	.939	.882	.960	.962	.946	.923
Unseen	.811	.655	.783	.709	.896	.867	.830	.755
Total	.895	.805	.885	.807	.938	.935	.906	.865

Table 6
Evaluation of the statistical word sense discrimination method over MAI1994$_{\text{RAND}}$ (**A** = accuracy, ***F*** = F-score)

Method	***Accuracy***		
	Monosemous	***Polysemous***	***Overall***
Pure rule-based	—	—	.946
Coll×Coll + rule-based	.849	.886	.852
Coll×Coll + first-sense	.849	.590	.826
Baseline	—	—	.760

Table 7
Token-level accuracy over MAI1994$_{\text{RAND}}$

We further converted the results described in Table 6 into a multiclass accuracy by calculating the proportion of JCVs which received the correct class assignation over all three interpretation types. That is, we combined the output of the three binary classifiers into a single multiclass, and calculated the ratio of JCVs for which the learned multiclass matched the gold-standard multiclass. This was found to be 0.819, which compares favourably with the 0.872 reported for the manual rule-based method of Uchiyama and Ishizaki (2003), as determined over a selection of 242 non-idiomatic JCV types taken from the Shin Meikai dictionary.

Next, we calculated the token-level classification accuracy over the MAI1994$_{\text{RAND}}$ dataset. Table 7 shows the accuracy of: (1) the pure rule-based system, (2) the statistical sense discrimination method coupled with rule-based disambiguation (**Coll×Coll + rule-based**), and (3) the purely statistical method (with first-sense disambiguation: **Coll×Coll + first-sense**); in the latter two cases, we break down the accuracy according to whether the JCV was classified as monosemous (in which case no form of disambiguation is employed) or polysemous. We also present a baseline based on the first-sense disambiguation heuristic independent of sense discrimination (i.e. we disambiguate each JCV occurrence according to the majority class for that V2 in the MAI1993 data).

All methods can be seen to significantly outperform the baseline. The margin between the pure rule-based method and the statistical method was a weighty 12 points (0.946 vs. 0.826). A small gain was achieved by coupling statistical sense discrimination with rule-based rather than first-sense disambiguation, but this effect was negligible in the overall performance due to the majority of JCVs being classified as monosemous. The lower-than-baseline performance of the statistical method over polysemous JCVs is due to the multiplicative effect of errors in word sense discrimination and disambiguation, but its impact on overall accuracy is relatively small due to the predominance of JCVs tagged as monosemous.

The disparity between the pure rule-based and statistical methods suggests that syntactic and semantic features have considerable potential to boost classification accuracy. Having said this, we must bear in mind that the rule-based method presupposes manual syntactic and semantic disambiguation of each JCV token instance. It would be quite possible to automate this process and reuse the existing rule set to perform JCV disambiguation, but it is by no means certain that a sufficiently accurate automatic parser and semantic tagger could be constructed which would enhance rather than diminish the overall accuracy. Perhaps a more realistic approach would be to complement the existing word sense discrimination technique with shallow contextual features as can be determined reliably from the output of a chunk parser, and attempt to improve our monosemous JCV accuracy.

The primary source of errors for the pure rule-based method is over-specialisation of rules, producing gaps. 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 and token not having been observed in Mai1993. For instance, Rule 4 in Section 4.1 would not generalise to inputs such as *Kare-wa wakamono-ni utai-kaketa* "He sang to the young people" [spatial] as *utau* "sing" is classified as a music and not action verb. In this case, *utai-kaketa* "sang to" would be misclassified according to the V2-default of aspectual. This motivates further investigation into the defeasibility/generalisation of conditions on certain rules.

6 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. It then employs a simple first-sense heuristic to disambiguate token occurrences of any JCV that is classified as polysemous. The rule-based method, on the other hand, relies on manual semantic analysis of the V1 and also argument structure in classifying JCVs at the token level. We established that both methods outperformed the first-sense baseline for the task, but that the rule-based method is superior to the statistical method (94.6% vs. 82.6%).

This suggests that the fully-automated statistical method provides a powerful means of token-level JCV sense-tagging, but that deep syntactical and semantic analysis can offer improvements in performance.

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.

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.

Acknowledgements

We would like to thank Francis Bond, Dan Flickinger, Ivan Sag, Koji Tsukamoto and the anonymous reviewers for their valuable input on this research. The second author was supported in part by NSF grant BCS-0094638 and also the Research Collaboration between NTT Communication Science Laboratories, Nippon Telegraph and Telephone Corporation and CSLI, Stanford University.

References

Baldwin, T. and F. Bond (2002). Multiword expressions: Some problems for Japanese NLP. In *Proc. of the 8th Annual Meeting of the Association for Natural Language Processing (Japan)*, Keihanna, Japan, pp. 379–82.

Baldwin, T. and A. Villavicencio (2002). Extracting the unextractable: A case study on verb-particles. In *Proc. of the 6th Conference on Natural Language Learning (CoNLL-2002)*, Taipei, Taiwan, pp. 98–104.

Bannard, C., T. Baldwin, and A. Lascarides (2003). A statistical approach to the semantics of verb-particles. In *Proc. of the ACL-2003 Workshop* *on Multiword Expressions: Analysis, Acquisition and Treatment*, Sapporo, Japan, pp. 65–72.

Daelemans, W., J. Zavrel, K. van der Sloot, and A. van den Bosch (2003). *TiMBL: Tilburg Memory Based Learner, version 5.0, Reference Guide.* ILK Technical Report 03-10.

Hashimoto, C. (2003). Japanese HPSG: Treatment of syntax and semantics of syntactically complex verbs. In *Information Processing Society of Japan SIG Notes.* (in Japanese).

Himeno, M. (2001). The nature of compound verbs. *Nihongogaku 240*(20), 6–15.

IPA (1987). *IPA Lexicon of the Japanese Language for Computers.* Tokyo, Japan. (In Japanese).

Kageyama, T. (1993). *Grammar and Word Formation.* Tokyo, Japan: Hitsuji Shobo.

Kageyama, T. (1999). Word formation. In *The Handbook of Japanese Linguistics*, pp. 297–325. Blackwell Publishers.

Kilgarriff, A. (2004). How dominant is the commonest sense of a word? Technical Report ITRI-04-10, Information Technology Research Institute, University of Brighton.

Kindaichi, K. (1999). *Shin Meikai Kokugo Dictionary* (5th ed.). Tokyo, Japan: Sanseido.

Lindner, S. (1983). *A Lexico-semantic Analysis of Verb-particle Constructions with Up and Out.* Indiana, USA: Indiana University Linguistics Club.

Mainichi Newspaper Co. (1993). *Mainichi Shimbun CD-ROM 1993.*

Mainichi Newspaper Co. (1994). *Mainichi Shimbun CD-ROM 1994.*

Martin, S. E. (1975). *A Reference Grammar of Japanese.* Rutland, USA and Tokyo, Japan: Tuttle.

Matsumoto, Y. (1996). *Complex Predicates in Japanese: A Syntactic and Semantic Study of the Notion 'Word'.* Stanford, USA: CSLI & Kurosio Publishers.

Matsumoto, Y. (1998). The combinatory possibilities in Japanese V-V lexical compounds. *Gengo Kenkyu 114*, 37–83.

Matsumoto, Y., A. Kitauchi, T. Yamashita, and Y. Hirano (2000). *Japanese Morphological Analysis System ChaSen Version 2.2.1 Manual.* Technical report, NAIST.

McCarthy, D., R. Koeling, J. Weeds, and J. Carroll (2004). Finding predominant senses in untagged text. In *Proc. of the 42nd Annual Meeting of the ACL*, Barcelona, Spain, pp. 280–7.

Niimi, K., Y. Yamaura, and T. Utsuno (1987). *Compound Verbs.* Tokyo, Japan: Aratake Shuppan. in Japanese.

Oono, S. and M. Hamanishi (1989). *Ruigo Shin Jiten.* Tokyo, Japan: Kadokawa Shoten.

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 International Conference on Intelligent Text Processing and Computational*

Linguistics (CICLing-2002), Mexico City, Mexico, pp. 1–15.

Schütze, H. (1998). Automatic word sense discrimination. *Computational Linguistics 24*(1), 97–123.

Shirai, S., Y. Ooyama, S. Takechi, K. Wakebe, and H. Aizawa (1998). Compiling Japanese and English corpus for compound verbs of Japanese origin. In *Proc. of the 57th Annual Meeting of IPSJ*, Nagoya, Japan, pp. 267–8. (in Japanese).

Tsujimura, N. (1996). *An Introduction to Japanese Linguistics.* Cambridge, USA: Blackwell.

Uchiyama, K. and T. Baldwin (2004). A machine learning method for disambiguating Japanese verb compounds. In *Proc. of the 10th Annual Meeting of the Association for Natural Language Processing (Japan)*, Tokyo, Japan. (in Japanese).

Uchiyama, K. and S. Ishizaki (2003). A disambiguation method for Japanese compound verbs. In *Proc. of the ACL-2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, Sapporo, Japan, pp. 81–8.

Villavicencio, A. and A. Copestake (2002). Verb-particle constructions in a computational grammar of English. In *Proc. of the 9th International Conference on Head-Driven Phrase Structure Grammar (HPSG-2002)*, Seoul, South Korea.