Paraphrasing of Japanese Light-verb Constructions Based on Lexical Conceptual Structure

Atsushi Fujita[†] **Kentaro Furihata**[†] **Kentaro Inui**[†] **Yuji Matsumoto**[†] **Koichi Takeuchi**[‡]
[†]Graduate School of Information Science,
Nara Institute of Science and Technology
{atsush-f,kenta-f,inui,matsu}@is.naist.jp
[‡]Department of Information Technology,
Okayama University
koichi@it.okayama-u.ac.jp

Abstract

Some particular classes of lexical paraphrases such as verb alteration and compound noun decomposition can be handled by a handful of general rules and lexical semantic knowledge. In this paper, we attempt to capture the regularity underlying these classes of paraphrases, focusing on the paraphrasing of Japanese light-verb constructions (LVCs). We propose a paraphrasing model for LVCs that is based on transforming the Lexical Conceptual Structures (LCSs) of verbal elements. We also propose a refinement of an existing LCS dictionary. Experimental results show that our LCS-based paraphrasing model characterizes some of the semantic features of those verbs required for generating paraphrases, such as the direction of an action and the relationship between arguments and surface cases.

1 Introduction

Automatic paraphrase generation technology offers the potential to bridge gaps between the authors and readers of documents. For example, a system that is capable of simplifying a given text, or showing the user several alternative expressions conveying the same content, would be useful for assisting a reader (Carroll et al., 1999; Inui et al., 2003).

In Japanese, like other languages, there are several classes of paraphrasing that exhibit a degree of regularity that allows them to be explained by a handful of sophisticated general rules and lexical semantic knowledge. For example, paraphrases associated with voice alteration, verb/case alteration, compounds, and lexical derivations all fall into such classes. In this paper, we focus our discussion on another useful class of paraphrases, namely, the paraphrasing of light-verb constructions (LVCs), and propose a computational model for generating paraphrases of this class.

Sentence (1s) is an example of an LVC[1]. An LVC is a verb phrase ("*kandou-o ataeta* (made an impression)" in (1s)) that consists of a light-verb ("*ataeta* (give-PAST)") that grammatically governs a nominalized verb ("*kandou* (an impression)") (also see Figure 1 in Section 2.2). A paraphrase of (1s) is sentence (1t), in which the nominalized verb functions as the main verb with its verbal form ("*kandou-s-as<u>e</u>-ta* (be impressed-CAU, PAST)").

(1) s. *Eiga-ga kare-ni kandou-o ataeta.*
film-NOM him-DAT impression-ACC give-PAST
The film made an impression on him.

t. *Eiga-ga kare-o kandou-s-ase-ta.*
film-NOM him-ACC be impressed-CAUSATIVE, PAST
The film impressed him.

To generate this type of paraphrase, we need a computational model that is capable of the following two classes of choice (also see Section 2.2):

Selection of the voice: The model needs to be able to choose the voice of the target sentence from active, passive, causative, etc. In example (1), the causative voice is chosen, which is indicated by the auxiliary verb "*ase* (causative)".

Reassignment of the cases: The model needs to be able to reassign a case marker to each argument of the main verb. In (1), the grammatical case of "*kare* (him)," which was originally assigned the dative case, is changed to accusative.

The task is not as simple as it may seem, because both decisions depend not only on the syntactic and semantic attributes of the light-verb, but also on those of the nominalized verb (Muraki, 1991).

In this paper, we propose a novel lexical semantics-based account of the LVC paraphrasing, which uses the theory of Lexical Conceptual Structure (LCS) of Japanese verbs (Kageyama, 1996; Takeuchi et al., 2001). The theory of LCS offers an advantage as the basis of lexical resources for paraphrasing, because it has been developed to explain varieties of linguistic phenomena including lexical derivations, the construction of compounds, and verb alteration (Levin, 1993; Dorr et al., 1995; Kageyama, 1996; Takeuchi et al., 2001), all of which are associated with the systematic paraphrasing we mentioned above.

The paraphrasing associated with LVCs is not idiosyncratic to Japanese but also appears commonly

[1]For each example, s denotes an input and t denotes its paraphrase.

in other languages such as English (Mel'čuk and Polguère, 1987; Iordanskaja et al., 1991; Dras, 1999, etc.), as indicated by the following examples.

(2) s. Steven *made an attempt* to stop playing.
t. Steven *attempted* to stop playing.

(3) s. It *had a noticeable effect* on the trade.
t. It *noticeably affected* the trade.

Our approach raises the interesting issue of whether the paraphrasing of LVCs can be modeled in an analogous way across languages.

Our aim in this paper are: (i) exploring the regularity of the LVC paraphrasing based a lexical semantics-based account, and (ii) assessing the immature Japanese semantic typology through a practical task.

The following sections describe our motivation, target, and related work on LVC paraphrasing (Section 2), the basics of LCS and the refinements we made (Section 3), our paraphrasing model (Section 4), and our experiments (Section 5). Finally, we conclude this paper with a brief of description of work to be done in the future (Section 6).

2 Motivation, target, and related work

2.1 Motivation

One of the critical issues that we face in paraphrase generation is how to develop and maintain knowledge resources that covers a sufficiently wide range of paraphrasing patterns such as those indicating that "to make an attempt" can be paraphrased into "to attempt," and that "potential" can be paraphrased into "possibility." Several attempts have been made to develop such resources manually (Sato, 1999; Dras, 1999; Inui and Nogami, 2001); those work have, however, tended to restrict their scope to specific classes of paraphrases, and cannot be used to construct a sufficiently comprehensive resource for practical applications.

There is another trend in the research in this field, namely, the automatic acquisition of paraphrase patterns from parallel or comparable corpora (Barzilay and McKeown, 2001; Lin and Pantel, 2001; Pang et al., 2003; Shinyama and Sekine, 2003, etc.). This type of approach may be able to reduce the cost of resource development. There are problems that must be overcome, however, before they can work practically. First, automatically acquired patterns tend to be complex. For example, from the paraphrase of (4s) into (4t), we can naively obtain the pattern: "*X* is purchased by *Y* ⇒ *Y* buys *X*."

(4) s. This car *was purchased by* him.
t. He *bought* this car.

This could also, however, be regarded as a combination of a simpler pattern of lexical paraphrasing ("purchase ⇒ buy") and a voice activization ("*X* be *VERB*-PP by *Y* ⇒ *Y VERB X*"). If we were to use an acquisition scheme that is not capable of decomposing such complex paraphrases correctly, we would have to collect a combinatorial number of paraphrases to gain the required coverage. Second, the results of automatic acquisition would likely include many inappropriate patterns, which would require manual correction. Manual correction, however, would be impractical if we were collecting a combinatorial number of patterns.

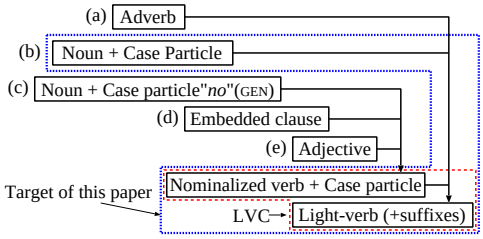

Figure 1: Dependency structure showing the range which the LVC paraphrasing affects.

Our approach to this dilemma is as follows: first, we manually develop the resources needed to cover those paraphrases that appear regularly, and then decompose and automatically refine the acquired paraphrasing patterns using those resources. The work reported in this paper is aimed at this resource development.

2.2 Target structure and required operations

Figure 1 shows the range which the LVC paraphrasing affects, where the solid boxes denote Japanese base-chunk so-called "*bunsetsu*."[2] Being involved in the paraphrasing, the modifiers of the LVC need the following operations:

Change of the dependence: The dependences of the elements (a) and (b) need to be changed because the original modifiee, the light-verb, is eliminated by the paraphrasing.

Re-conjugation: The conjugation form of the elements (d), (e), and occasionally (c) need to be changed according to the category change of their modifiee, the nominalized verb.

Reassignment of the cases: As described in the previous section, the case markers of the elements (b) and often (c) need to be reassigned.

Selection of the voice: The voice of the nominalized verb needs to be chosen according to the combination of the nominalized verb, the light-verb, and the original voice.

The first two operations are trivial in the field of text generation. Moreover, they can be done independently of the LVC paraphrasing. The most delicate operation is for the element (c) because it acts either as an adverb or as a case, relying on the con-

[2]The modifiee of the LVC is not affected because the part-of-speech of the light-verb and main verb are the same.

Table 1: Examples of LCS

Verb	LCS for verb	Verb phrase
move	[y MOVE TO z]	My sister (Theme) moves to a neighboring town (Goal).
transmit	[x CONTROL [y MOVE TO z]]	The enzyme (Agent) transmits messages (Theme) to the muscles (Goal).
locate	[y BE AT z]	The school (Theme) locates near the river (Goal).
maintain	[x CONTROL [y BE AT z]]	He (Agent) maintains a machine (Theme) in good condition (Goal).

text. In the former case, it needs the second operation. In the latter case, it needs the third operation as well as the element (b).

In this paper, we take into account only the element (b), namely, the sibling cases of the nominalized verb.

2.3 Related work

Based on the Meaning-Text Theory (Mel'čuk and Polguère, 1987), Iordanskaja et al. (1991) proposes a set of paraphrasing rules including one for LVC paraphrasing. Their rule heavily relies on what are called lexical functions, by which they virtually specify all the choices relevant to LVC paraphrasing for every combination of nominalized verb and light-verb individually. Our approach is to employ lexical semantics to provide a general account of those classes of choices.

On the other hand, Kaji and Kurohashi (2004) proposes a paraphrasing model which bases on an ordinary dictionary. Given an input LVC, their model paraphrases it using the gloss of both the nominalized verb and the light-verb with the semantic feature of the light-verb. Their model looks robust because of the availability of an ordinary dictionary. However, their model fails to explain the difference in the voice selection between examples (5) and (6) since it selects the voice based only on the light-verb — in their approach, the light-verb "*ukeru* (to receive)" always maps to the passive voice irrespective of the nominalized verb.

(5) s. *Enkai-eno shoutai-o uketa*.
party-GEN invitation-ACC receive-PAST
I received an invitation to the party.
t. *Enkai-ni shoutai-s-are-ta*.
party-DAT invite-PAS, PAST
I was invited to the party.

(6) s. *Kare-no hanashi-ni*
his-GEN talk-DAT
kandou-o uketa.
impression-ACC receive-PAST
I was given a good impression by his talk.
t. *Kare-no hanashi-ni kandou-shi-ta*.
his-GEN talk-DAT be impressed-ACT, PAST
I was impressed by his talk.

In (Kaji and Kurohashi, 2004), the target expression is restricted only to the LVC itself (also see Figure 1). Hence, their model is unable to reassign the cases as we saw in example (1).

3 Lexical Conceptual Structure

3.1 Basic framework of LCS

The theory of Lexical Conceptual Structure (LCS) associates a verb with a semantic structure as exemplified by Table 1. An LCS consists of semantic predicates ("CONTROL," "BE AT," etc.) and their argument slots (x, y, z). Argument slots x, y, and z correspond to the semantic roles "Agent," "Theme," and "Goal," respectively. Taking the LCS of the verb "transmit" as an example, [y MOVE TO z] denotes the state of affairs that the state of the "Theme" changes to the "Goal," and [x CONTROL . . .] denotes that the "Agent" causes the state change.

3.2 Refinements

We make use of the TLCS dictionary, a Japanese verb LCS dictionary developed by Takeuchi et al. (2001), because it offers the following advantages:

- It is based on solid linguistic work, as in (Kageyama, 1996).
- Its scale is considerably larger than any other existing collections of verb LCS entries.
- It provides a set of concrete rules for LCS assignment, which ensures the reliability of the dictionary.

In spite of these advantages, our preliminary examination of the dictionary revealed that further refinements were needed. To refine the typology of TLCS, we collected the following sets of words:

Nominalized verbs: We regard "*sahen*-nouns"[3] and nominal forms of verbs as nominalized verbs. We retrieved 1,210 nominalized verbs from the TLCS dictionary.

Light-verbs: Since a verb takes different meanings when it is a part of LVCs with different case particles, we collected pairs $\langle c, v \rangle$ of case particle c and verb v in the following way:

Step 1. We collected 876,101 types of triplets $\langle n, c, v \rangle$ of nominalized verb n, case particle c, and base form of verb v from the parsed[4] sentences of newspaper articles[5].

[3] A *sahen*-noun is a verbal noun in Japanese, which acts as a verb in the form of "*sahen*-noun + *suru*".

[4] We used the statistical Japanese dependency parser CaboCha (Kudo and Matsumoto, 2002) for parsing. http://chasen.naist.jp/~taku/software/cabocha/

[5] Excerpts from 9 years of the Mainichi Shinbun and 10 years of the Nihon Keizai Shinbun, giving a total of 25,061,504 sentences, were used.

Table 2: Extensions of LCS

	Verb	Verb phrase and its LCS representation
Ext.1	*hankou-suru*	[[*Ken*]y BE AGAINST [*parents*]z]
	(resist)	*Ken-ga oya-ni hankou-suru.*
		Ken-NOM parents-DAT resist-PRES (Ken resists his parents.)
Ext.2	*ukeru*	[BECOME [[*salesclerk*]z BE WITH [[*complaint*]y MOVE FROM [*customer*]x TO [*salesclerk*]z]]]
	(receive)	*Ten'in-ga kyaku-kara kujo-o ukeru.*
		salesclerk-NOM customer-ABL complaint-ACC receive-PRES
		(The salesclerk receives a complaint from a customer.)
Ext.3	*motomeru*	[[*Ken*]x CONTROL [[*apology*]y MOVE FROM [*George*]z TO [FILLED]]][6]
	(ask)	*Ken-ga George-ni shazai-o motomeru.*
		Ken-NOM George-DAT apology-ACC ask-PRES (Ken asks George for an apology.)
Ext.4	*kandou-suru*	[BECOME [[*Ken*]z BE WITH [[FILLED]y MOVE FROM [*music*]x TO [*Ken*]z]]]
	(be impressed)	*Ken-ga ongaku-ni kandou-suru.*
		Ken-NOM music-DAT be impressed-PRES (Ken is impressed by the music.)

Step 2. For each of the 50 most frequent $\langle c, v\rangle$ tuples, we extracted the 10 most frequent $\langle n, c, v\rangle$.

Step 3. Each $\langle n, c, v\rangle$ was manually evaluated to determine whether it was an LVC. If any of 10 triplets was determined to be an LVC, $\langle c, v\rangle$ was merged into the list of light-verbs. As a result, we collected 40 types of $\langle c, v\rangle$ for light-verbs.

Through investigating the above 1,210 nominalized verbs and 40 light-verbs, we extended the typology of TLCS as shown below (also see Table 2).

Ext. 1. Treatment of "Partner": The dative case of "*hankou-suru* (resist)" and "*eikyo-suru* (affect)" does not indicate the "Goal" of the action but the "Partner."

Ext. 2. Verbs of obtaining (Levin, 1993): In contrast with "*ataeru* (give)," the nominative case of "*ukeru* (receive)" and "*eru* (acquire)" is the "Goal" of the "Theme," while the ablative case indicates "Source."

Ext. 3. Require verb: "*motomeru* (ask)" and "*yokyu-suru* (require)" denote the existence of the external "Agent" who controls the action of the other "Agent" or "Theme."

Ext. 4. Verbs of psychological state (Levin, 1993): "*kandou-suru* (be impressed)" and "*osoreru* (fear)" indicate the change of psychological state of the "Agent." The ascriptive part of the change has to be described.

Consequently, we defined a new LCS typology consisting of 16 types. Note that more than one LCS can be assigned to a verb if it has a polysemy. For convenience, we refer to the extended dictionary as the LCSdic[7].

[6]The predicate "FILLED" represents an implicit argument of the verb and the verb assigned this LCS cannot take this argument. In the LCS of the verb "sign", for example, "FILLED" in [x CONTROL [BECOME [[FILLED]y BE AT z]]] denotes the name of "Agent."

[7]The latest version of the LCSdic is available from http://cl.it.okayama-u.ac.jp/rsc/lcs/

4 Paraphrasing model

In this section, we describe how we generate paraphrases of LVCs. Figure 2 illustrates how our model paraphrases the LVC of example (7).

(7) s. *Ken-ga eiga-ni shigeki-o uketa.*
Ken-NOM film-DAT inspiration-ACC receive-PAST
Ken received an inspiration from the film.

t. *Ken-ga eiga-ni shigeki-s-are-ta.*
Ken-NOM film-DAT inspire-PAS, PAST
Ken was inspired by the film.

The idea is to exploit the LCS representation as a semantic representation and to model the LVC paraphrasing by the transformation of the LCS representation. The process consists of the following three steps:

Step 1. Semantic analysis: The model first analyzes a given input sentence including an LVC to obtain its semantic structure in terms of the LCS representation. In Figure 2, this step produces LCS_{V1}.

Step 2. Semantic transformation (LCS transformation): The model then transfers the obtained semantic structure to another semantic structure so that the target structure consists of the LCS of the nominalized verb of the input. In our example, this step generates LCS_{N1} together with the supplement [BECOME [...]].

Step 3. Surface generation: Having obtained the target LCS representation, the model finally lexicalizes it to generate the output sentence.

So, the key issue is how to control the second step, namely, the transformation of the LCS representation.

The rest of this section elaborates on each step, using different symbols to denote arguments; x, y, and z for LCS_V, and x', y', and z' for LCS_N.

4.1 Semantic analysis

Given an input sentence, which we assume to be a simple clause with an LVC, we first look up the LCS template LCS_{V0} for the given light-verb, and then apply the *case assignment rule*, below (Takeuchi et

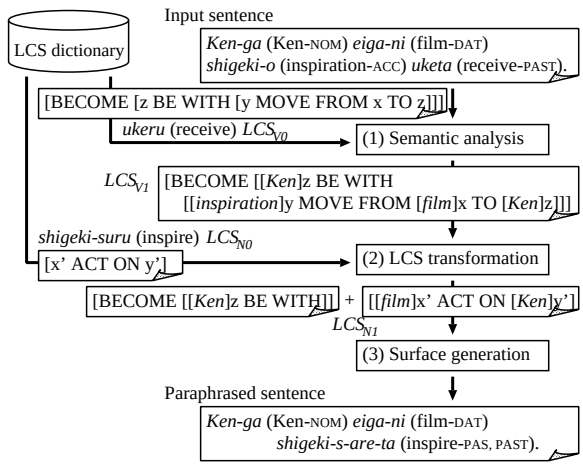

Figure 2: The LCS-based paraphrasing model.

al., 2001), to obtain its LCS representation LCS_{V1}:

Case assignment rule:

- In the case of the LCS_{V0} having argument x, fill the leftmost argument of the LCS_{V0} with the nominative case of the input, the second leftmost with the accusative, and the rest with the dative.
- Otherwise, fill arguments y and z of the LCS_{V0} with the nominative and the dative, respectively.

In the example shown in Figure 2, the nominative "*Ken*" fills the leftmost argument z. Accordingly, the accusative "*shigeki* (inspiration)" and the dative "*eiga* (film)" fill y and x, respectively.

(8) s. *Ken-ga eiga-ni shigeki-o uketa*.
Ken-NOM film-DAT inspiration-ACC receive-PAST
Ken received an inspiration from the film.
LCS_{V0} [BECOME [z BE WITH [y MOVE FROM x TO z]]]
LCS_{V1} [BECOME [[*Ken*]z BE WITH [[*inspiration*]y MOVE FROM [*film*]x TO [*Ken*]z]]]

4.2 LCS transformation

The second step of our paraphrasing model matches the resultant LCS representation (LCS_{V1} in Figure 2) with the LCS of the nominalized verb (LCS_{N0}) to generate the target LCS representation (LCS_{N1}). Figure 3 shows a more detailed view of this process for the example shown in Figure 2.

4.2.1 Predicate matching

The first step is to determine the predicate in LCS_{V1} that should be matched with the predicate in LCS_{N0}. Assuming that only the agentivity is relevant to the selection of the voice in the paraphrasing of LVC, which is our primary concern, we classify the semantic predicates into the following two classes:

Agentive predicates: "CONTROL," "ACT ON," "ACT," "BE AGAINST," and "MOVE FROM TO."

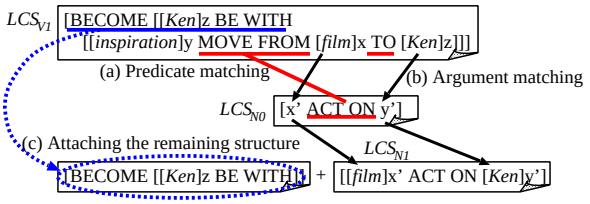

Figure 3: LCS transformation.

State of affair predicates: "MOVE TO," "BE AT," and "BE WITH."

Aspectual predicates: "BECOME."

We also assume that any pair of predicates of the same class is allowed to match, and that the aspectual predicates are ignored. In our example, "MOVE FROM TO" matches "ACT ON," as shown in Figure 3.

LCS representations have right-branching (or right-embedding) structures. Since inner-embedded predicates denote the state of affairs, they take priority in the matching. In other words, the matching proceeds from the rightmost inner predicates to the outer predicates.

Having matched the predicates, we then fill each argument slot in LCS_{N0} with its corresponding argument in LCS_{V1}. In Figure 3, argument z is matched with y', and x with x'. As a result, "*Ken*" comes to the y' slot and "*eiga* (film)" comes to the x' slot[8].

This process is repeated until the leftmost predicate in LCS_{N0} or that in LCS_{V1} is matched.

4.2.2 Treatment of non-transfered predicates

If LCS_{V1} has any non-transfered predicates when the predicate matching has been completed, they represent the semantic content that is not covered by LCS_{N1} and which needs to be lexicalized by auxiliary linguistic devices such as voice auxiliaries. In the case of Figure 3, [BECOME [[*Ken*]z BE WITH]] in LCS_{V1} remains non-transfered. In such a case, we attach the non-transfered predicates to LCS_{N0}, which are then lexicalized by auxiliaries in the next step, the surface generation.

4.3 Surface generation

We again apply the aforementioned case assignment rule to generate a sentence from the resultant LCS representation. In this process, the model makes the final decisions on the selection of the voice and the reassignment of the cases, according to the following decision list:

[8] When an argument is filled with another LCS, arguments within the inner LCS are also matched. Likewise, with regard to an assumption that the input sentences are periphrastic, we introduced some exceptional rules. That is, arguments filled with the implicit filler represented by "FILLED" or the target nominalized verb N are never matched, and "Goal" in LCS_{V1} can be matched to "Theme" in LCS_{N0}.

1. If the attached predicate is filled with the same argument as the leftmost argument in LCS_{N1}, the "active" voice is selected and the case structure is left as is.
2. If the argument of the attached predicate has the same value as either z' or y' in LCS_{N1}, lexicalization is performed to make the argument a subject. Therefore, the "passive" voice is selected and case alternation (passivization) is applied.
3. If the attached predicate is "BE WITH" and its argument has the same value as x' in LCS_{N1}, the "causative" voice is selected and case alternation (causativizaton) is applied.
4. If the attached predicate is an agentive predicate, and its argument is filled with a value different from those of the other arguments, then the "causative" voice is selected and case alternation (causativization) is applied.
5. Otherwise, no modification is applied.

Since the example in Figure 2 satisfies the second condition, the model chooses "*s-are-ru* (passive)" and passivizes the sentence so that "*Ken*" fills the nominative case.

(9) LCS_{N1} [BECOME [[*Ken*]z BE WITH]]
+ [[*film*]x' ACT ON [*Ken*]y']
t. *Ken-ga eiga-ni shigeki-s-are-ta.*
Ken-NOM film-DAT inspire-PAS, PAST
Ken was inspired by the film.

5 Experiment

5.1 Paraphrase generation and evaluation

To empirically evaluate our paraphrasing model and the LCSdic, and to clarify the remaining problems, we analyzed a set of automatically generated paraphrase candidates. The sentences used in the experiment were collected in the following way:

Step 1. From the 876,101 types of triplet $\langle n, c, v\rangle$ collected in Section 3.2, 23,608 types of $\langle n, c, v\rangle$ were extracted, whose components, n and $\langle c, v\rangle$, are listed in the LCSdic.

Step 2. For each of the 245 most frequent $\langle n, c, v\rangle$, the 3 most frequent simple clauses including the $\langle n, c, v\rangle$ were extracted from the corpus from which $\langle n, c, v\rangle$ s were extracted in Section 3.2. As a result, we collected 735 sentences.

Step 3. We input these 735 sentences into our paraphrasing model, and then automatically generated paraphrase candidates. When more than one LCS is assigned to a verb in the LCSdic due to its polysemy or ergative verb such as "*kaifuku-suru* (recover)," our model generates all the possible paraphrase candidates. As a result, 825 paraphrase candidates, that is, at least one for each input, were generated.

Table 3: Error sources

Correct candidates	**621** (75.8%)
Erroneous candidates	**198** (24.2%)
Definition of LCS	**30**
LCS for light-verb	24
LCS for nominalized verb	6
Paraphrasing model	**61**
LCS transformation algorithm	59
Treatment of "*suru* (to do)"	2
Ambiguity	**107**
Ambiguous thematic role of dative	78
Recognition of LVC	24
Selection of transitive/intransitive	5

We manually classified the resultant 825 paraphrase candidates into 621 correct and 198 erroneous candidates. The remaining 6 candidates were not classified. The precision of the paraphrase generation was 75.8% (621 / 819).

5.2 Error analysis

To clarify the cause of the erroneous paraphrases, we manually classified 198 erroneous paraphrase candidates. Table 3 lists the error sources.

5.2.1 LCS transformation algorithm

The experiment came close to confirming that the right-first matching algorithm in our paraphrasing model operates correctly. Unfortunately, the matching rules produced some erroneous paraphrases in LCS transformation.

Errors in predicate matching: To paraphrase (10s) below, "CONTROL" in LCS_{V1} must be matched with "CONTROL" in LCS_{N0}, and x to x'. However, our model first matched "CONTROL" in LCS_{V1} with "MOVE FROM TO" in LCS_{N0}. Thus, x was incorrectly matched with z' and x' remained empty. The desired form of LCS_{N1} is shown in (11).

(10) s. *kacho-ga buka-ni*
section-chief-NOM subordinate-DAT
shiji-o dasu.
order-ACC issue-PRES
The section chief issues orders to his subordinates.
(N="order", V="issue")
LCS_{V1} [[*chief*]x CONTROL [BECOME [[*order*]y BE AT [*subordinate*]z]]]
LCS_{N0} [x' CONTROL [y' MOVE FROM z' TO [FILLED]]]
LCS_{N1}*[x' CONTROL [[*subordinate*]y' MOVE FROM [*chief*]z' TO [FILLED]]]

(11) LCS_{N1} [[*chief*]x' CONTROL [y' MOVE FROM [*subordinate*] TO [FILLED]]]

This error was caused by the mis-matching of "CONTROL" with "MOVE FROM TO." Although we regard some predicates as being in the same classes as those described in Section 4.2.1, these need to be considered carefully. In particular

"MOVE FROM TO" needs further investigation because it causes many errors whenever it has the "FILLED" argument.

Errors in argument matching: Even if all the predicates are matched properly, there would still be a chance of errors being caused by incorrect argument matching. With the present algorithm, z can be matched with y' if and only if z' contains "FILLED." In the case of (12), however, z has to be matched with y', even though z' is empty. The desired form of LCS_{N1} is shown in (13).

(12) s. *Jikan-ni seigen-ga aru.*
time-DAT limitation-NOM exist-PRES
There is a time limitation.
(N="limitation", V="exist")
LCS_{V1} [[*limitation*]y BE AT [*time*]z]
LCS_{N0} [x' CONTROL [BECOME [y' BE AT z']]]
LCS_{N1}*[x' CONTROL [BECOME [y' BE AT [*time*]z']]]

(13) LCS_{N1} [x' CONTROL [BECOME [[*timey*]y' BE AT z']]]

5.2.2 Ambiguous thematic role of dative

In contrast to dative cases in English, in Japanese, the dative case has ambiguity. That is, it can be a complement to the verb or an adjunct[9]. However, since LCS is not capable of determining whether the case is a complement or an adjunct, z is occasionally incorrectly filled with an adjunct. For example, "*medo-ni*" in (14s) should not fill z, because it acts as an adverb, even though it consists of a noun, "*medo* (prospect)" and a case particle for the dative. We found that 78 erroneous candidates constitute this most dominant type of errors.

(14) s. *Kin'you-o medo-ni sagyo-o susumeru.*
Friday-NOM by-DAT work-ACC carry on-PRES
I plan to finish the work by Friday.
(N="work", V="carry")
LCS_{V0} [x CONTROL [BECOME [y BE AT z]]]
LCS_{V1}*[x CONTROL [BECOME [[*work*]y BE AT [*by*]z]]]

The ambiguity of dative cases in Japanese has been discussed in the literature of linguistics and some natural language processing tasks (Muraki, 1991). To date, however, a practical compliment/adjunct classifier has not been established. We plan to address this topic in our future research. Preliminary investigation revealed that only certain groups of nouns can constitute both compliments and adjuncts according to the governing verb. Therefore, generally whether a word acts as a complement is determined without combining it with the verb.

[9](Muraki, 1991) classifies dative cases into 11 thematic roles that can be regarded as complements. In contrast, there is no typology of dative cases that act as adjuncts.

5.2.3 Recognition of LVC

In our model, we assume that a triplet $\langle n, c, v \rangle$ consisting of a nominalized verb n and a light-verb tuple $\langle c, v \rangle$ from our vocabulary lists (see Section 3.2) always act as an LVC. However, not only the triplet itself but also its context sometimes affects whether the given triplet can be paraphrased. For example, we regard "*imi-ga aru*" as an LVC, because the nominalized verb "*imi*" and the tuple $\langle$"*ga*", "*aru*"$\rangle$ appear in the vocabulary lists. However, the $\langle n, c, v \rangle$ in (15s) does not act as an LVC, while the same triplet in (16s) does.

(15) s. *Sanka-suru-koto-ni imi-ga aru.*
to participate-DAT meaning-NOM exist-PRES
There is meaning in participating.
t.*Sanka-suru-koto-o imi-suru.*
to participate-ACC mean-ACT, PRES
*It means to participate in it.

(16) s. *"kennel"-niwa inugoya-toiu*
"kennel"-TOP doghouse-OF
imi-ga aru.
meaning-NOM exist-PRES
"kennel" has the meaning of doghouse.
t. *"kennel"-wa inugoya-o imi-suru.*
"kennel"-TOP doghouse-ACC mean-ACT, PRES
"kennel" means doghouse.

The above difference is caused by the polysemy of the nominalized verb "*imi*" that denotes "worth" in the context of (15s), but "meaning" in (16s). Although incorporating word sense disambiguation using contextual clues complicates our model, in fact only a limited number of nominalized verbs are polysemous. We therefore expect that we can list them up and use this as a trigger for making a decision as to whether we need to take the context into account. Namely, given a $\langle n, c, v \rangle$, we would be able to classify it into (a) a main verb phrase, (b) a delicate case in terms of the dependence of its context, and (c) an LVC.

We can adopt a different approach to avoiding incorrect paraphrase generation. As described in Section 5.1, our model generates all the possible paraphrase candidates when more than one LCS is assigned to a verb. Similarly, our approach can be extended to (i) over-generate paraphrase candidates by considering the polysemy of not only assigned LCS types, but also that of nominalized verbs (see (15s) and (16s)) and whether the given $\langle n, c, v \rangle$ is an LVC, and (ii) revise or reject the incorrect candidates by using handcrafted solid rules or statistical language models.

6 Conclusion and future work

In this paper, we presented an LCS-based paraphrasing model for LVCs and an extension of an ex-

isting LCS dictionary. Our model achieved an accuracy of 75.8% in selecting the voice and reassigning the cases.

To make our paraphrasing model more accurate, further analysis is needed, especially for the LCS transformation stage described in Section 4.2. Similarly, several levels of disambiguation should also be solved. The Japanese LCS typology has to be refined from the theoretical point of view. For example, since extensions are no more than human intuition, we must discuss how we can assign LCSs for given verbs based on explicit language tests, as described in (Takeuchi et al., 2001).

In future research, we will also extend our LCS-based approach to other classes of paraphrases that exhibit some regularity, such as verb alteration and compound noun decomposition as shown in (17) and (18), below. LCS has been discussed as a means of explaining the difference between transitive/intransitive verbs, and the construction of compounds. Therefore, our next goal is to show the applicability of LCS through practical tasks, namely, paraphrasing.

(17) s. *Jishin-ga building-o kowashita.*
earthquake-NOM building-DAT destroy-PAST
The earthquake destroyed the building.
t. *Jishin-de building-ga kowareta.*
earthquake-LOC building-NOM be destroyed-PAST
The building was destroyed in the earthquake.

(18) s. *Kare-wa kikai sousa-ga jouzu-da.*
he-TOP machine- operation-NOM good-COPULA
He is good at operating the machine.
t. *Kare-wa kikai-o jouzu-ni sousa-suru.*
he-TOP machine-DAT well-ADV operate-PRES
He operates machines well.

References

R. Barzilay and K. R. McKeown. 2001. Extracting paraphrases from a parallel corpus. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics* (*ACL*), pages 50–57.

J. Carroll, G. Minnen, D. Pearce, Y. Canning, S. Devlin, and J. Tait. 1999. Simplifying text for language-impaired readers. In *Proceedings of the 9th Conference of the European Chapter of the Association for Computational Linguistics* (*EACL*), pages 269–270.

B. J. Dorr, J. Garman, and A. Weinberg. 1995. From syntactic encodings to thematic roles: building lexical entries for interlingual MT. *Machine Translation*, 9(3):71–100.

M. Dras. 1999. *Tree adjoining grammar and the reluctant paraphrasing of text*. Ph.D. thesis, Department of Computing, Macquarie University.

K. Inui and M. Nogami. 2001. A paraphrase-based exploration of cohesiveness criteria. In *Proceedings of the 8th European Workshop on Natulal Language Generation* (*EWNLG*), pages 101–110.

K. Inui, A. Fujita, T. Takahashi, R. Iida, and T. Iwakura. 2003. Text simplification for reading assistance: a project note. In *Proceedings of the 2nd International Workshop on Paraphrasing: Paraphrase Acquisition and Applications* (*IWP*), pages 9–16.

L. Iordanskaja, R. Kittredge, and A. Polguère. 1991. Lexical selection and paraphrase in a meaning-text generation model. In *Paris et al.* (*Eds.*) *Natural Language Generation in Artificial Intelligence and Computational Linguistics*, pages 293–312. Kluwer Academic Publishers.

T. Kageyama, editor. 1996. *Verb semantics*. Kuroshio Publishers. (in Japanese).

N. Kaji and S. Kurohashi. 2004. Recognition and paraphrasing of periphrastic and overlapping verb phrases. In *Proceedings of the 4th International Conference on Language Resources and Evaluation* (*LREC*) *Workshop on Methodologies and Evaluation of Multiword Units in Real-world Application*.

T. Kudo and Y. Matsumoto. 2002. Japanese dependency analysis using cascaded chunking. In *Proceedings of 6th Conference on Natural Language Learning* (*CoNLL*), pages 63–69.

B. Levin. 1993. *English verb classes and alternations: a preliminary investigation*. Chicago Press.

D. Lin and P. Pantel. 2001. Discovery of inference rules for question answering. *Natural Language Engineering*, 7(4):343–360.

I. Mel'čuk and A. Polguère. 1987. A formal lexicon in meaning-text theory (or how to do lexica with words). *Computational Linguistics*, 13(3-4):261–275.

S. Muraki. 1991. *Various aspects of Japanese verbs*. Hitsuji Syobo. (in Japanese).

B. Pang, K. Knight, and D. Marcu. 2003. Syntax-based alignment of multiple translations: extracting paraphrases and generating new sentences. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics* (*HLT-NAACL*), pages 102–109.

S. Sato. 1999. Automatic paraphrase of technical papers' titles. *Journal of Information Processing Society of Japan*, 40(7):2937–2945. (in Japanese).

Y. Shinyama and S. Sekine. 2003. Paraphrase acquisition for information extraction. In *Proceedings of the 2nd International Workshop on Paraphrasing: Paraphrase Acquisition and Applications* (*IWP*), pages 65–71.

K. Takeuchi, K. Uchiyama, S. Yoshioka, K. Kageura, and T. Koyama. 2001. Categorising deverbal nouns based on lexical conceptual structure for analysing Japanese compounds. In *Proceedings of IEEE System, Man, and Cybernetics Conference*, pages 904–909.