

Verbal Case Frame Acquisition from Bilingual Corpora*

Takehito Utsuro[†]

[†]Dept. of Electrical Engineering
Kyoto University
Yoshida-honmachi, Sakyo-Ku,
Kyoto, 606, Japan

Yuji Matsumoto[‡]

Makoto Nagao[‡]

[‡]Graduate School of Information Science
Advanced Institute of Science and Technology, Nara
8916-5, Takayama-cho, Ikoma-shi,
Nara, 630-01, Japan

Abstract

This paper describes a method for acquiring surface case frames of Japanese verbs from bilingual corpora. Our approach makes use of translation examples in two distinct languages that have quite different syntactic structures and word meanings (such as English and Japanese). In many cases, both syntactic and semantic ambiguities are resolved by comparing analyzed results of sentences in both languages, and it becomes possible to extract lexical knowledge with much less human interaction compared with monolingual approaches. The interaction is limited to the crucial points for discriminating multiple senses of verbal case frames that the system detects automatically. The system reduces human interaction by virtue of English information, which works as the heuristics for the discrimination.

1 Introduction

For a practical research in natural language processing, it is indispensable to develop a large scale semantic dictionary. It is especially important to devise techniques of compiling semantic dictionaries from natural language texts in existing human dictionaries and/or large corpora. There are, however, at least two difficulties in analyzing real texts:

1. The problem of syntactic ambiguities
2. The problem of polysemy

Our approach to solving these difficulties is to make use of translation examples of two distinct languages that have quite different syntactic structures and word meanings (such as English and Japanese), and to compare analyzed results of these languages [Utsuro *et al.*,

1992]. In many cases, translation pairs of those languages have different types of syntactic ambiguities, and comparison of syntactic structures of both languages helps to resolve them. Multiple senses of a word may be reduced by the information of translated word of the opposite language [Dagan *et al.*, 1991; Tokunaga and Tanaka, 1991]. For example, in the case of the translation example given in Example 1, both syntactic and semantic ambiguities are resolved.

Example 1

I hung my coat on the hook.

私 は 上着 を かぎ に かけた。
I TOP coat ACC hook on hung

1. Syntactic disambiguation

The English sentence of Example 1 is syntactically ambiguous because the prepositional phrase “on the hook” can modify either of the verb “hung” and the noun phrase “my coat”. On the other hand, in the Japanese sentence, the phrase “かぎ, に” can modify nothing but the verb “かけた”. Thus, if knowledge about word equivalence pairs such as {*I*, 私}, {*hung*, かけた}, {*coat*, 上着}, {*hook*, かぎ} are available from a bilingual dictionary, the ambiguity of pp-attachment is resolved by matching the structures of the two sentences.

2. Semantic disambiguation

The verb “かける” in the Japanese sentence is a typical Japanese polysemy. This verb has six sub-entries in a Japanese dictionary that has about 70,000 entries, and ten translatable English verbs (“hang”, “spend”, “play”, etc.) in a Japanese-English dictionary. It is not easy to associate the surface word “かける” with its exact sense. With the translation example, the corresponding English verb “hang” helps to find the sense of “かける”.

From this viewpoint, we proposed the framework of *Lexical Knowledge Acquisition* [Utsuro *et al.*, 1992], as in Figure 1. In our framework, pairs of English and Japanese sentences are syntactically analyzed without any semantic knowledge and are translated into (disjunctive) feature descriptions, which represent dependency structures of the sentences. Then these feature descriptions of the two languages are compared, or unified, using knowledge of word equivalence obtained from a bilingual dictionary. In the process of unification, (some of)

*The authors would like to thank the editorial staff of Kodansha for permission to use the data of Japanese-English dictionary, and also thank Dr. Shouichi Yokoyama of ETL, and Prof. Hozumi Tanaka and Dr. Takenobu Tokunaga of Tokyo Institute of Technology, for providing them the data of Japanese-English dictionary. This work is partly supported by the Grants from Ministry of Education, #03245103.

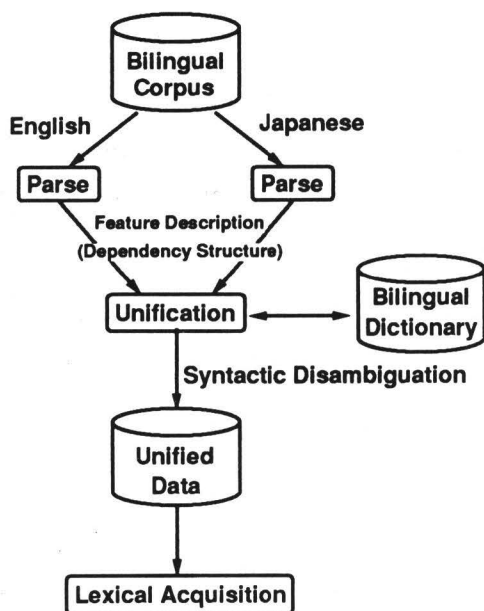


Figure 1: The Framework of *Lexical Knowledge Acquisition from Bilingual Corpora*

syntactic ambiguities are resolved¹. In the case of the translation example of Example 1, the bilingually unified feature description (henceforth *bilingual feature description*) is obtained unambiguously as follows:

$$\left[\begin{array}{l} \text{pred : } \langle \text{hang, かける} \rangle \\ \text{tense : } \text{past} \\ \langle \text{subj, は} \rangle : \left[\text{pred : } \langle \text{I, 私} \rangle \right] \\ \langle \text{obj, を} \rangle : \left[\begin{array}{l} \text{pred : } \langle \text{coat, 上着} \rangle \\ \text{spec : } \text{my} \end{array} \right] \\ \langle \text{on, に} \rangle : \left[\begin{array}{l} \text{pred : } \langle \text{hook, かぎ} \rangle \\ \text{spec : } \text{the} \end{array} \right] \end{array} \right]$$

In this *bilingual feature description*, the verb is represented as the pair of English and Japanese atomic values $\langle \text{hang, かける} \rangle$, and the feature labels are represented as pairs of English and Japanese feature labels like $\langle \text{subj, は} \rangle$, $\langle \text{obj, を} \rangle$, $\langle \text{on, に} \rangle$. This is also the case with the nouns of the case elements. From the collection of these disambiguated data, lexical knowledge such as surface case frames of verbs is acquired.

The bilingual corpus we have collected includes 40,000 translation examples obtained from a machine readable Japanese-English dictionary (Kodansha Japanese-English Dictionary [Shimizu and Narita, 1979]). In [Utsuro *et al.*, 1992], we described the method of resolving the syntactic ambiguities of translation examples by unifying the feature descriptions of the two languages. Although the success ratio of unambiguous analysis of Japanese sentences alone are 22%, the success ratio of

acquiring unambiguous case structure of each sentence goes up to 59% when the bilingual unification is employed. In this paper, we describe the process of acquiring surface case frames of Japanese verbs in detail.

When designing a framework of knowledge acquisition by computer program, we have to decide how much human-machine interaction is necessary. There are several research activities of compiling semantic dictionaries or knowledge bases by hand [EDR, 1989; IPA, 1987; Lenat and Guha, 1990]. Several difficulties arise in this approach. Huge amount of manual labor is necessary for compiling large scale dictionaries, and the expansion of the dictionaries is difficult once they are compiled by hand. Moreover, the results tend to be unstable because the quality of human writers is not uniform. On the other hand, there are other research activities that attempt to compile semantic dictionaries automatically or semiautomatically by extracting lexical knowledge from natural language texts taken from human dictionaries or large corpora (e.g. [Brent, 1991; Tomiura *et al.*, 1992; Tsurumaru *et al.*, 1991]). Computer program may do away with most of the above difficulties of scale and stability. However, currently available knowledge resources are too poor to make it possible to extract lexical knowledge from natural language texts in a fully automatic way. Thus, we consider that human interaction is necessary to some extent though it should be kept to a minimum in order to make the results stable. We take an approach that the system asks questions to human only when it comes to a crucial point of acquiring lexical knowledge. In order to acquire stable lexical knowledge, we limit the interaction between human and the system to a selection among a few choices.

This paper introduces a method of acquiring surface case frames of verbs from the collection of *bilingual feature descriptions*. The acquisition process is depicted in Figure 2. We start with a set of *bilingual feature descriptions* which contain the target verb. The acquisition process is divided into two sub-processes:

1. Discrimination of multiple senses of the target verb.
2. Construction of the case frame for each sense of the target verb, making use of a nominal thesaurus.

For discriminating multiple senses of a verb, the system detects the clue to the discrimination using some heuristics, and asks the human instructor whether the detected clue is correct or not. When acquiring case frames of Japanese verbs, our heuristics for the clue detection becomes much more reliable compared with monolingual approaches, since Japanese words are paired with English information such as the corresponding English verb and English case labels. From the results of the heuristics and human interaction, the set of *bilingual feature descriptions* is divided into subsets according to the different senses or usages of the target verb. Then, from each subset, a surface case frame of the target verb is constructed.

In Chapter 2, we introduce the data structure of the nominal thesaurus and the *bilingual feature descriptions*. In Chapter 3, we discuss how we can discriminate multiple senses of a Japanese verb. Chapter 4 describes the

¹Recently we developed another method to find structural matching between parallel sentences of two languages, which is presented in [Matsumoto *et al.*, 1993]. The main idea is to find a set of pairs of minimal corresponding substructures of the dependency structures between parallel sentences.

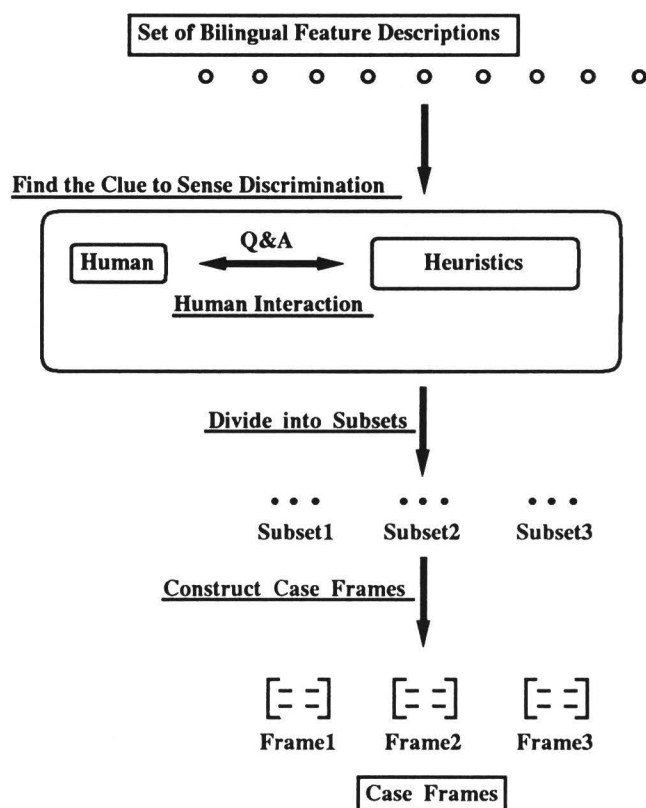


Figure 2: Process of Acquiring Case Frames of A Verb

process of constructing a case frame for each sense of a verb. The results of some experiments and evaluation are shown in Chapter 5.

2 Data Structure

2.1 Nominal Thesaurus

We use semantic categories in a nominal thesaurus for describing the semantic restrictions of case slots. A nominal thesaurus is regarded as a tree in which each node represents a semantic category. We define a nominal thesaurus as a finite partially ordered set $[SC, \preceq]$, where SC is the set of the semantic categories in the thesaurus, and $\preceq$ is the superordinate-subordinate relation of semantic categories. $[SC, \preceq]$ has a single greatest element T which corresponds to the root of the tree. $a \vee b$ represents the least upper bound of two semantic categories a and b . We assume that a unique least upper bound exists for all a and b .

At present we use an on-line thesaurus called 'Bunrui Goi Hyou'(BGH)[NLRI, 1964]. BGH has a six-layered abstraction hierarchy and more than 60,000 Japanese words are assigned at the leaves. In its nominal part, about 45,000 words are assigned at the leaves, and five semantic categories are allocated at the next level from T , which are *abstract relations*, *agents of human activities*, *human activities*, *products* and *natural objects and natural phenomena*. One word is assigned under several leaf semantic categories in the thesaurus when it has sev-

eral senses. This thesaurus is not reliable enough for our target of acquiring case frames of verbs. It is, however, the most precise and broad covering Japanese nominal thesaurus obtainable to us currently.

2.2 Bilingual Feature Description

We acquire surface case frames of verbs from a collection of *bilingual feature descriptions* [Utsuro *et al.*, 1992]. A *bilingual feature description* is a feature description where the atoms and the labels are expanded to include paired symbols consisting of bilingual vocabularies:

$$\left[\begin{array}{l} pred : \langle V_E, V_J \rangle \\ voice_e : Voice_E \\ aux_J : Aux_J \\ \langle l_{E1}, p_{J1} \rangle : \left[\begin{array}{l} pred : \langle N_{E1}, N_{J1} \rangle \\ sem : SEM_1 \end{array} \right] \\ \vdots \\ \langle l_{En}, p_{Jn} \rangle : \left[\begin{array}{l} pred : \langle N_{En}, N_{Jn} \rangle \\ sem : SEM_n \end{array} \right] \end{array} \right]$$

V_J and V_E indicate the corresponding Japanese and English verbs, $Voice_E$ denotes the voice of the English sentence, and Aux_J denotes the auxiliary verb of the Japanese sentence. $p_{J1}, \dots, p_{Jn}$ indicate the Japanese case markers and $l_{E1}, \dots, l_{En}$ indicate the corresponding English case markers. $N_{J1}, \dots, N_{Jn}$ indicate Japanese nouns of the case elements, and $N_{E1}, \dots, N_{En}$ indicate the corresponding English nouns. When a Japanese noun N_{Ji} has several senses, it appears at several leaf semantic categories in the thesaurus. SEM_i is a set of those semantic categories, and is referred to as a *semantic label*.

$$SEM_i = \{Sem_1, \dots, Sem_k\}$$

When describing the semantic restrictions of case slots, this semantic ambiguity must be resolved and the correct semantic category Sem_j must be selected.

3 Discriminating Multiple Senses of A Verb

This chapter first discusses what can be the clues to discriminating multiple senses of a Japanese verb, then proposes a method for discriminating multiple senses of a verb.

3.1 Clues to Discriminating Multiple Senses of A Verb

One verb may have several senses and therefore may have as many case frames. In our framework, when discriminating multiple senses of a Japanese verb, English information, such as the corresponding English verb and English case labels, as well as the *semantic labels* of Japanese nouns in case slots, are available. The clues are obtained by making use of such information.

Difference of the Corresponding English Verbs

When one Japanese verb has several senses, distinct senses are often translated into different English verbs. In such a case, the different English verbs can be the clue to discriminating senses of the Japanese verb. For

example, Japanese verb “かける” has several senses and it is translated into different English verbs.

- かぎ に 上着 を かける
hook on coat ACC hang
/ hang my coat on the hook
- 服 に 金 を かける
clothes on money ACC spend
/ spend money on clothes
- 音楽 を かける / play some music
music ACC play

Difference of the Corresponding English Case Labels

When one Japanese verb has several case frames, it often happens that some case marker commonly occurring in distinct case frames play different roles. For example, Japanese verb “書く” has two case frames according to the different usages. Both case frames have a case marked by postpositional particle “に”.

- 黒板 に 字 を 書く
blackboard on character ACC write
/ write a character on the blackboard
- 友 に 手紙 を 書く
friend to letter ACC write
/ write a letter to a friend

These “に” cases have different senses and the difference of them can be the clue to discriminating these case frames of “書く”. As observed from this example, these “に” cases are translated into different English prepositions, “on” and “to”. Such an evidence appearing in English translation is available in our framework as the clue to discriminating these case frames.

Semantic Labels of Nouns in Case Slots

Sometimes, the corresponding English information is of no use to discriminate multiple senses of a Japanese verb. Although a Japanese verb has several case frames, it may be translated into the same English verb regardless of the difference of its senses. Also, it may happen that a Japanese case marker which appear in different case frames is translated into the same English case label, even when it plays a distinct role in each case frame. The following are two examples of case frames of a Japanese verb “上げる”.

- 帽子 を 上げる / raise one's hat
hat ACC raise
- 叫び声 を 上げる / raise a cry
cry ACC raise

Both of them have the case marked by postpositional particle “を”, and their corresponding English case label is “obj” for both of them. These case frames have no difference in their syntactic level. However, the nouns in these case slots, “帽子”(hat) and “叫び声”(cry), have completely disjoint *semantic labels*. So the disjointness of *semantic labels* of these nouns is useful as the clue to discriminating these case frames. Disjointness of *semantic labels* are calculated after generalizing those *semantic labels* to a predetermined level in the thesaurus. In the case of BGH, currently the next level from T is regarded

as that predetermined level, which consists of five categories.

3.2 Method for Discriminating Multiple Senses of A Verb

When acquiring case frames of a Japanese verb V_J , we start with a set S of *bilingual feature descriptions* that have V_J at ‘pred’ label. The problem of discriminating multiple senses of one Japanese verb V_J is regarded as the problem of discriminating two *bilingual feature descriptions* in which the corresponding senses of the verb V_J are different from each other. To illustrate the method for discriminating multiple senses of the verb V_J , we define the *case description* (CD) in the form of the feature description:

$$\left[\begin{array}{l} \text{pred} : \langle V_E, V_J \rangle \\ \langle l_E, p_J \rangle : [\text{sem} : SEM] \end{array} \right]$$

A *case description* is a part of a *bilingual feature description* that includes the information necessary for the subsequent acquiring process. For each pair of feature labels $\langle l_E, p_J \rangle$ in a *bilingual feature description*, a *case description* is extracted.

As we have shown in the previous section, the clues to discriminating two *bilingual feature descriptions* are found in the corresponding English verb V_E , English case labels $l_{E1}, \dots, l_{Em}$ and *semantic labels* of the Japanese nouns in case slots. When extracting a *case description* from a *bilingual feature description*, the *semantic label* of the case slot is generalized to the predetermined level in the thesaurus. In the case of BGH, the predetermined level is the next level from T. This means that the *semantic label SEM* in the *case description* is a subset of the set of five topmost semantic categories.

By comparing *case descriptions*, discrimination of case frames of *bilingual feature descriptions* is done through the following *heuristics* and *human interaction*. Then the set S of *bilingual feature descriptions* is divided into subsets $S_1, \dots, S_l$.

Heuristics

The system finds the pair of *case descriptions* which can be the clue to discriminating multiple senses of the Japanese verb V_J by some *heuristics*. We take into consideration two *case descriptions* of Japanese verb V_J that have the same Japanese case marker p_J . The clue is found from the *case descriptions* based on the discussion in the previous section, which are expressed in terms of *case descriptions* as follows.

- (a) Two *case descriptions* CD_1 and CD_2 have different English verbs.

$$CD_1 = \left[\begin{array}{l} \text{pred} : \langle V_{E1}, V_J \rangle \\ \langle l_{E1}, p_J \rangle : [\text{sem} : SEM_1] \end{array} \right]$$

$$CD_2 = \left[\begin{array}{l} \text{pred} : \langle V_{E2}, V_J \rangle \\ \langle l_{E2}, p_J \rangle : [\text{sem} : SEM_2] \end{array} \right]$$

- (b) CD_1 and CD_2 have the same English verb but have different English case labels for the Japanese case marker p_J .

$$CD_1 = \left[\begin{array}{l} \text{pred} : \langle V_E, V_J \rangle \\ \langle \underline{l}_{E1}, p_J \rangle : [\text{sem} : SEM_1] \end{array} \right]$$

$$CD_2 = \left[\begin{array}{l} \text{pred} : \langle V_E, V_J \rangle \\ \langle l_{E2}, p_J \rangle : [\text{sem} : SEM_2] \end{array} \right]$$

(c) CD_1 and CD_2 have the same English verb and the same English case label for p_J , but their *semantic labels* in case slots are mutually disjoint. In the case of BGH, this difference of *semantic labels* is detected among the five topmost semantic categories.

$$CD_1 = \left[\begin{array}{l} \text{pred} : \langle V_E, V_J \rangle \\ \langle l_E, p_J \rangle : [\text{sem} : SEM_1] \end{array} \right]$$

$$CD_2 = \left[\begin{array}{l} \text{pred} : \langle V_E, V_J \rangle \\ \langle l_E, p_J \rangle : [\text{sem} : SEM_2] \end{array} \right]$$

Human Interaction

After the system finds a pair of *case description* CD_1 and CD_2 that can be the clue, it asks the human instructor if the pair is really the clue. The question is presented with CD_1 , CD_2 and a few sample sentences involving *bilingual feature descriptions*. The human instructor selects one of the following three answers.

Ans1: Clue

CD_1 and CD_2 represent different cases and they must be included in different case frames of the Japanese verb V_J .

Ans2: Co-occur

Although CD_1 and CD_2 represent different cases, they can co-occur in a sentence.

For example, a Japanese verb “書く” has two cases both of which are marked by a postpositional particle “で” and are translated into different English preposition “in” and “with”.

- 英語 で 書く / write in English
English in write
- 筆 で 書く / write with brush
brush with write

These two cases are different cases of “書く” and can co-occur in a sentence. Thus they cannot be the clue to discriminating the case frames.

Ans3: Equivalent

Although CD_1 and CD_2 are detected as the clue, they should be considered as equivalent cases. For example, the two case frames of the Japanese verb “見る” below should be equivalent although they have different English verbs.

- 映画 を 見る / see a movie
movie ACC see
- 試合 を 見る / watch a game
game ACC watch

In *heuristics* (a) ~ (c), the clues are found and asked to the human instructor in the order of (c), (b), (a). This is because the clues are more likely to become *equivalent* in this order and consequently this order reduces the total number of questions.

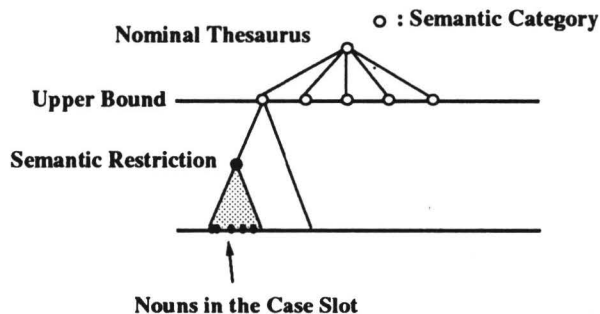


Figure 3: Calculating the Semantic Restriction of A Case Slot

Dividing the Set of the Bilingual Feature Descriptions into Distinct Subsets

From the results of the *human interaction*, some *case descriptions* are judged as equivalent. With this equivalence relation, the set of all *case descriptions* is divided into equivalent classes, $ECD_1, \dots, ECD_p$. Besides, from the results of the *human interaction*, a set of the clues has been obtained. A clue is represented as a pair of *case descriptions*, $\langle CD_1, CD_2 \rangle$. A clue can be extended to a pair of *ECDs*, $\langle ECD_1, ECD_2 \rangle$, where ECD_1 includes CD_1 and ECD_2 includes CD_2 . Then the set of the clues is represented as a set of pairs of *ECDs*.

An expanded clue $\langle ECD_1, ECD_2 \rangle$ defines an incompatible relation between *bilingual feature descriptions*. The set S of all *bilingual feature descriptions* is then divided into maximal compatible subsets $S_1, \dots, S_l$ making use of this incompatible relation².

4 Constructing Case Frames of A Verb

After S is divided into $S_1, \dots, S_l$, the system constructs a case frame Fr_i for each S_i . In general, a surface case frame of a Japanese verb V_J is represented as the following feature description:

$$\left[\begin{array}{l} \text{pred} : V_J \\ p_{J1} : [\text{sem} : SR_1] \\ \vdots \\ p_{Jm} : [\text{sem} : SR_m] \end{array} \right]$$

$SR_1, \dots, SR_m$ represent the semantic restrictions of the case slots. Each SR_j represents the possible semantic categories of the case slot and is a set of the semantic categories of the thesaurus. However, in this section, we represent the case frame Fr_i as a set of $\langle ECD, SR \rangle$, where ECD is an equivalent class of *case descriptions* which appeared in any of the *bilingual feature descriptions* in S_i and SR is the semantic restriction, that is,

$$Fr_i = \{ \langle ECD_1, SR_1 \rangle, \dots, \langle ECD_n, SR_n \rangle \}$$

Each SR_j is calculated from $SEM_1, \dots, SEM_N$, which are the *semantic labels* of all case elements of p_j contained in S_i , where p_j is the case marker which appears

²Though this division is not always unique, *bilingual feature descriptions* that belong to more than one compatible sets are regarded as ambiguous and are neglected.

Table 1: Collected Case Descriptions of “買う”

	English Verb and Case Labels	Semantic Label	Freq.	Examples
1	buy-(subj, が)	HUMAN	6	私 (I)
2	buy-(obj, を)	PRODUCTS	13	車 (car)
		NATURAL	3	土地 (land)
3	incur-(subj, が)	HUMAN	2	彼 (he)
4	incur-(obj, を)	HUMAN-ACTIVITIES	3	反感 (antipathy)
5	appreciate-(obj, を)	HUMAN-ACTIVITIES	3	努力 (effort)

in ECD_j .³ Each SEM_r ($1 \leq r \leq N$) is a set of the semantic categories of the nominal thesaurus. We assume that SR_j is represented by semantic categories less general than a predetermined upper bound, as described in Figure 3. In the case of BGH, currently the upper bound is the five topmost categories. SR_j is meant to be the least general semantic category that covers at least one semantic category in every SEM_r ($1 \leq r \leq N$). If the least general semantic category is not determined uniquely, SR_j becomes a set of the least semantic categories.

5 Experiment and Evaluation

We have obtained *bilingual feature descriptions* of 16 verbs and acquired surface case frames from them. The following example shows the process of acquiring case frames of a Japanese verb “買う” from 48 *bilingual feature descriptions*. Case descriptions of “買う” collected from the set of *bilingual feature descriptions* are shown in Table 1. Among them, the system finds clues to discriminating multiple senses of “買う” by heuristics, and asks human if they are appropriate as in Table 2. In the case of “買う”, the English verbs “buy”, “incur” and “appreciate” become the clue to sense discrimination. An example of questions and answers is like below.

Question:

What is the relation between these two cases ?

1: buy-(obj, を)

E.g.: 私が家を買う/I buy a house.

彼が土地を買う/He buys a land.

2: incur-(obj, を)

E.g.: 私が反感を買う/I incur antipathy.

彼が恨みを買う/He incurs enmity.

Choices

(Ans1) Clue (Ans2) Co-occur (Ans3) Equivalent

Answer: ⇒ (Ans1) Clue

Then three case frames of “買う” are acquired as in Table 3. (The semantic restriction is described by the five topmost semantic categories in BGH for simplicity.) For the comparison of the acquired case frames with those of hand-made case frames, we show the case frames of “買う” in IPAL dictionary (*IPA Lexicon of the Japanese Language for computers*) [IPA, 1987] in Table 4. IPAL has case frames of 861 basic Japanese verbs and uses about 20 semantic markers to describe semantic restrictions. “買う” has four case frames in IPAL. Case frame

³Note that all case descriptions in an *ECD* have a common Japanese case marker.

Table 2: Questions and Answers about the Clue to Sense Discrimination of “買う”

Pairs of Case Descriptions Detected by Heuristics		Answer
buy-(obj, を)-PRODUCTS	buy-(obj, を)-NATURAL	Equivalent
buy-(obj, を)	incur-(obj, を)	Clue
buy-(obj, を)	appreciate-(obj, を)	Clue
incur-(obj, を)	appreciate-(obj, を)	Clue

Table 3: Acquired Case Frames of “買う”

	Case	Semantic Restriction	Examples of Slot Element
1	buy-(subj, が)	HUMAN	私 (I)
	buy-(obj, を)	PRODUCTS, NATURAL	家 (house), 土地 (land)
2	incur-(subj, が)	HUMAN	彼 (he)
	incur-(obj, を)	HUMAN-ACTIVITIES	反感 (antipathy)
3	appreciate-(obj, を)	HUMAN-ACTIVITIES	努力 (effort)

1 and 2 represent the sense of “buy” and their difference is the “に (for)” case. In our corpus, the “に (for)” case does not appear. Although some cases in case frames of IPAL do not appear in the acquired case frames, case frames are discriminated almost in the same way.

In Table 5, results of experiments for 16 Japanese verbs are shown. Numbers in the column A and I are those of acquired case frames and case frames in IPAL. $X \wedge \neg Y$ represents the number of case frames in X that are not contained in Y . $X > Y$ represents that several case frames in X correspond to one case frame in Y . In this column, $n-1$ means that n case frames correspond to one case frame. “Eg.” is the number of examples, i.e., the number of *bilingual feature descriptions* used to acquire case frames. For some verbs, there are case frames which exist in IPAL but not acquired. This is because our corpus does not contain enough examples. (The poorest one is “かける”, for which 28 case frames couldn’t be acquired.) On the other hand, among acquired case frames, 12 case frames don’t exist in IPAL. In IPAL, idiomatic usages of verbs don’t have their own entries and are included in the entries of literal usages. In our experiment, idiomatic usages are discriminated from literal ones and have their own entries. 6 out of 12 case frames are idiomatic usages.

Some acquired case frames correspond to one case

Table 4: Case Frames of “買う” in IPAL

	Case Label	Semantic Marker
1, 2 (buy)	が (agent)	HUMAN
	を (object)	CONCRETE, ABSTRACT
	に (for)	HUMAN) only in case frame 2
3 (incur)	が (agent)	HUMAN
	を (object)	MENTAL
4 (appreciate)	が (agent)	HUMAN
	を (object)	CHARACTER, MENTAL, ACTION

Table 5: Comparing Acquired Case Frames with IPAL

Verbs	X > Y: Several Case Frames in X Correspond to One Case Frame in Y							Eg.
	A	I	I $\wedge$ $\neg A$	A $\wedge$ $\neg I$	A > I	A < I		
受ける (receive)	5	9	6	0	3-1	0		88
うつ (hit)	3	23	20	0	0	0		37
買う (buy)	3	4	0	0	0	1-2		48
かかる (cost)	3	29	25	0	0	1-2		35
書く (write)	3	2	0	0	2-1	0		115
かける (hang)	7	35	28	1	0	1-2		59
聞く (hear)	3	6	2	0	2-1	1-2		85
来る (come)	3	12	8	0	0	1-2		52
知る (know)	1	5	2	0	0	1-3		69
つく (stick)	12	31	23	4	0	0		83
つける (attach)	10	14	6	1	2-1	0		108
出る (go out)	10	32	23	1	0	0		89
とる (take)	10	29	18	2	0	1-6, (1-2) ^{x3}		118
ひく (pull)	7	14	8	2	0	1-2		45
見る (see)	1	13	10	0	0	1-3		106
持つ (have)	5	12	4	1	0	1-2, (1-3) ^{x2}		69

frame in IPAL. Observed from the results of these examples, it looks desirable to discriminate some of case frames in IPAL into finer-grained ones. For example, for the case frames of “つける”, IPAL contains “刃をつける”(put an edge on a sword) and “理由をつける”(give some reason) in one case frame which has the meaning of “add or attach something to something”. In our experiments, these two usages are discriminated. On the other hand, some case frames in IPAL correspond to one acquired case frame. The reason is that the system couldn't find the clue to discriminating multiple senses of a verb by the heuristics. This is partly caused by the nominal thesaurus we are using. For example, it is hard for the system to discriminate the two examples, “噂 (HUMAN ACTIVITIES) を知る (know the rumor)” and “人生の苦勞 (HUMAN ACTIVITIES) を知る (know the bitterness of life)”. “知る” means “acquire some information or knowledge from some source” in the former and “understand the essence of things” in the latter. But the object nouns in both sentences belong to the semantic category HUMAN ACTIVITIES. And also for the verbs “とる” and “持つ”, it happens that some case frames in IPAL correspond to one acquired case frame much more often than other verbs. This is because the Japanese verb can be translated into one English verb in several senses (e.g. “take” for “とる” and “have” for “持つ”). In these cases, it is hard for the system to find effective clues to sense discrimination of a verb.

Although some difficulties might exist, the results suggest that it is quite possible to acquire case frames of verbs from bilingual corpora if enough translation examples and a suitable nominal thesaurus are available. We think it possible to construct a large scale bilingual corpus in the near future.

6 Concluding Remarks

We described a method of acquiring surface case frames of Japanese verbs from bilingual corpora. In the method, English information such as translated English verbs and

case labels are indispensably useful when discriminating multiple senses of a verb. The results of experiments show that it is quite possible to acquire case frames of verbs from bilingual corpora.

If the corpus contains some irregular sentences such as metaphoric expressions or erroneous sentences, acquired case frames may become incorrect. We guess that some metaphoric expressions are extracted as distinct case frames. But, it will be shown by our future work. It is necessary to develop techniques for acquiring the most probable case frames from noisy examples. The experiments has shown that much larger bilingual corpus is necessary for case frame acquisition of enough number of verbs. We have collected translated newspaper editorial articles of about 21,000 sentences and are currently developing a robust system to analyze them and to extract corresponding verb phrases from which case frames are acquired. We believe that the proposed method is also applicable to acquiring case frames of English verbs and translation patterns of verbs between two languages.

References

- [Brent, 1991] M. R. Brent. Automatic acquisition of subcategorization frames from untagged text. In *Proceedings of the 29th Annual Meeting of ACL*, pages 209-214, 1991.
- [Dagan et al., 1991] I. Dagan, A. Itai, and U. Schwall. Two languages are more informative than one. In *Proceedings of the 29th Annual Meeting of ACL*, pages 130-137, 1991.
- [EDR, 1989] EDR. (Japan Electric Dictionary Research Institute, Ltd.). *Conceptual Dictionary, 2nd. Edition*, TR-012, 1989. (in Japanese).
- [IPA, 1987] IPA. (Information-technology Promotion Agency, Japan). *IPA Lexicon of the Japanese Language for computers IPAL (Basic Verbs)*, 1987. (in Japanese).
- [Lenat and Guha, 1990] D. B. Lenat and R. V. Guha. *Building Large Knowledge-based Systems*. Addison-Wesley, 1990.
- [Matsumoto et al., 1993] Y. Matsumoto, H. Ishimoto, T. Utsuro, and M. Nagao. Structural matching of bilingual texts. In *Proceedings of the 31th Annual Meeting of ACL*, 1993.
- [NLRI, 1964] NLRI. (National Language Research Institute). *Word List by Semantic Principles*. Syuei Syuppan, 1964. (in Japanese).
- [Shimizu and Narita, 1979] M. Shimizu and S. Narita, editors. *Japanese-English Dictionary*. Kodansha Gakujutsu Bunko, 1979.
- [Tokunaga and Tanaka, 1991] T. Tokunaga and H. Tanaka. The automatic extraction of conceptual items from bilingual dictionaries. *Journal of Japan Society for Artificial Intelligence*, 6(2):228-235, 1991. (in Japanese).
- [Tomiura et al., 1992] Y. Tomiura, T. Nakamura, et al. Logical form of hierarchical relation on verbs and extracting it from definition sentences in a Japanese dictionary. In *Proceedings of the 14th COLING*, pages 574-580, 1992.
- [Tsurumaru et al., 1991] H. Tsurumaru, K. Takesita, et al. An approach to thesaurus construction. *IPSJ-WGNL*, 91(37 (91-NL-83)), 1991. (in Japanese).
- [Utsuro et al., 1992] T. Utsuro, Y. Matsumoto, and M. Nagao. Lexical knowledge acquisition from bilingual corpora. In *Proceedings of the 14th COLING*, pages 581-587, 1992.