# Lexical Knowledge Acquisition from Bilingual Corpora

Takehito UTSURO* Yuji MATSUMOTO Makoto NAGAO

Dept. of Electrical Engineering, Kyoto University
Yoshida-honmachi, Sakyo-Ku, Kyoto, 606, Japan
utsuro@kuee.kyoto-u.ac.jp

## Abstract

For practical research in natural language processing, it is indispensable to develop a large scale semantic dictionary for computers. It is especially important to improve the techniques for compiling semantic dictionaries from natural language texts such as those in existing human dictionaries or in large corpora. However, there are at least two difficulties in analyzing existing texts: the problem of syntactic ambiguities and the problem of polysemy. Our approach to solve these difficulties is to make use of translation examples in two distinct languages that have quite different syntactic structures and word meanings. The reason we took this approach is that in many cases both syntactic and semantic ambiguities are resolved by comparing analyzed results from both languages. In this paper, we propose a method for resolving the syntactic ambiguities of translation examples of bilingual corpora and a method for acquiring lexical knowledge, such as case frames of verbs and attribute sets of nouns.

## 1 Introduction

It has become widely accepted that developing a large scale semantic dictionary is indispensable to future natural language research. In recent years, several research activities for compiling semantic dictionaries for natural language processing have been undertaken. One of the approaches in this research is attempts to compile dictionaries by hand. Japan Electronic Dictionary Research Institute (EDR) is now compiling conceptual dictionaries[5] by hand with the help of software tools. Information-technology Promotion Agency (IPA), Japan, has also compiled *IPA Lexicon of the Japanese Language for computers* (IPAL)[4]. IPAL has 861 entries for basic Japanese verbs. Cyc project attempts to assemble a massive knowledge base covering human common-sense knowledge[7]. However, this approach suffers from problems such as a huge amount of manual labor, difficulties in extending the dictionaries, unstable results, and so forth.

Another approach is to compile dictionaries using some techniques of lexical knowledge acquisition. One such approach is to extract hierarchical relations or a thesaurus of conceptual items from human dictionaries in an automatic way. Tsurumaru et al. studied to construct a thesaurus of nominal concepts from noun definitions[13]. Tomiura et al. also extracted superordinate-subordinate relation between verbs from the defining sentences in IPAL[12]. Besides these researches, there are other several research activities for lexical knowledge acquisition, which syntactically analyze the sentences in large corpora and attempt to extract lexical knowledge from statistical data [3] [1]. Most of the works undertake shallow analysis of texts and they extract only superficial lexical information.

For the development of the techniques of knowledge acquisition from natural language texts, it is very important to improve the latter approach of compiling semantic dictionaries by computer programs. However, there are at least two basic difficulties in this approach.

1. **The problem of syntactic ambiguities**

   When analyzing a sentence, syntactic ambiguities often remain. So it is not easy to obtain correct parsed results automatically.

2. **The problem of polysemy**

   It often happens that one word has several meanings and corresponds to several concepts. So it is not easy to associate one surface word with one correct conceptual item.

Our approach to solve these difficulties is to make use of translation examples in two distinct languages that have quite different syntactic structures and word meanings (such as English and Japanese), and to compare analyzed results from each language. In many cases, the two languages have different types of syntactic ambiguities, and comparison of syntactic structures of both languages helps to resolve the ambiguities. Also, a pair of bilingually equivalent surface words helps to associate the words with conceptual

*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, ETL, and Prof. Hozumi TANAKA and Dr. Takenobu TOKUNAGA, Tokyo Institute of Technology, for providing us the data of Japanese-English dictionary. This work is partly supported by the Grants from Ministry of Education, #03245103.

words helps to associate the words with conceptual items, because the intersection of conceptual items that each surface word has could be considered as one conceptual item[11] [2]. For example, in the case of the translation example given in Example 1, both syntactic and semantic ambiguities are resolved.

**Example 1**

E: I hung my coat on the hook.
J: 私 (I) は (topic) 上着 (coat) を (case-marker) かぎ (hook) に (case-marker) かけた (hung)。

1. **Syntactic disambiguation**

   The English sentence in Example 1 is syntactically ambiguous because the prepositional phrase *"on the hook"* can modify both the verb *"hung"* and the noun phrase *"my coat"* using grammatical knowledge only. 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 bilingual dictionaries, the ambiguity of pp-attachment is resolved by syntactically 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 subentries in a Japanese dictionary that has about 70,000 entries, and ten English equivalent verbs (*"hang"*, *"spend"*, *"play"*, etc.) in a Japanese-English dictionary that has about 50,000 entries. So, it is not easy to associate the surface word "かける" with its exact meaning. However, with the translation example, the corresponding English verb such as *"hang"* helps to find the meaning of the Japanese verb "かける".

In this paper, we propose a method for resolving the syntactic ambiguities of translation examples in bilingual corpora and a method for acquiring lexical knowledge, such as case frames of verbs and attribute sets of nouns. In our framework, first a pair of sentences of both languages are syntactically analyzed[1] and translated into feature descriptions, which represent dependency structures of the phrases in the sentences. Although feature descriptions are generated by grammatical knowledge only, they are quite suitable to represent case frames of verbs. Then these feature descriptions of the two languages are compared, or unified, using knowledge about word equivalence from bilingual dictionaries. In this matching process, one word in the English sentence could be equivalent to several words in the translated Japanese sentence. Also one word in the Japanese sentence could be equivalent to several words in the translated English sentence. In order to realize the matching process between two languages including these several word equivalence cases, we introduce a unification algorithm based on sets of compatible pairs of atomic values and feature labels in Chapter 2.

In Chapter 3, we statistically evaluated the process of syntactic disambiguation. The success ratio of disambiguation is about 63~68 % for translation examples in a Japanese-English dictionary. At present, we have already collected about 50,000 translation examples from a machine readable Japanese-English dictionary (Kodansha Japanese-English Dictionary [10]) and an English learners' textbook. We have extracted case frames for several verbs as a simple experiment. The results are described in Chapter 4.

[1] The Japanese morphological analyzer has 14 part of speech and about 36,000 words. The English dictionary contains about 55,000 words. The current Japanese and English grammars consist of 85 DCG rules and 135 DCG rules.

# 2 Unification of Feature Descriptions of Two Languages

## 2.1 Unification based on Sets of Compatible Pairs of Features and Values

In our framework of sentence analysis, a sentence in each language is parsed and translated into feature descriptions, which represent dependency structures of the phrases in the sentence. In this section, we basically use and extend Kasper and Rounds' notation of *feature description logic* (FDL [6]) to describe our unification algorithm of feature descriptions, except that we don't use path equivalence.

When unifying feature descriptions of two languages, knowledge about word equivalence taken from bilingual dictionaries is used to decide whether an atomic value of one language is compatible with an atomic value of the other language. This is also the case with feature labels. Knowledge about word equivalence from bilingual dictionaries can be regarded as knowledge about compatibility of atomic values and feature labels of feature descriptions. From this standpoint, we introduce a unification algorithm based on sets of compatible pairs of atomic values and feature labels.

**Data Structure**

Let $A$ and $L$ be sets of symbols used to denote atomic values and feature labels. Let $C_A$ and $C_L$ be sets of compatible pairs of atomic values and feature labels. That is, $C_A$ is the set of pairs of atomic values such as $\langle a_i, a_j \rangle (a_i, a_j \in A)$, where $a_i$ and $a_j$ are consistent and unifiable, and $C_L$ is the set of pairs of feature labels like $\langle l_i, l_j \rangle (l_i, l_j \in L)$, where $l_i$ and $l_j$ are consistent

and unifiable[2,3].

The syntax for formulas of the *FDL with Sets of Compatible Pairs* (FDLC) is given below.

| | |
|---|---|
| $NIL$ | denoting no information |
| $TOP$ | denoting inconsistent information |
| $a$ | where $a \in A$, to describe atomic values |
| $\langle a_i, a_j \rangle$ | where $a_i, a_j \in A$ and $\langle a_i, a_j \rangle \in C_A$, to describe pairs of atomic values |
| $l : \phi$ | where $l \in L$ and $\phi \in$ FDLC, to describe structures in which the feature labeled by $l$ has a value described by $\phi$ |
| $\langle l_i, l_j \rangle : \phi$ | where $l_i, l_j \in L$ and $\langle l_i, l_j \rangle \in C_L$ and $\phi \in$ FDLC, to describe structures in which the feature labeled by $\langle l_i, l_j \rangle$ has a value described by $\phi$ |
| $\phi \wedge \psi$ | where $\phi, \psi \in$ FDLC |

### Unification Algorithm

Because of the compatibility sets, there is not necessarily a unique most general unifier of two feature descriptions. When applying this algorithm to unify feature descriptions between two languages, we collect all possible unified feature descriptions and find the most overlapping unifier by a scoring function, which is introduced later. The following definition of UNIFY returns one possible unified feature description. We collect all possible unified feature descriptions.

> Function UNIFY$(f, g)$ returns one possible unified *feature description*:
> where $f$ and $g$ are *feature descriptions*.

1. If $f = NIL$, then return $g$
2. Else if $g = NIL$, then return $f$
3. Else if $f = TOP$ or $g = TOP$,
   then return $TOP$
4. Else if $f, g \in A \cup C_A$ and $f = g$
   then return $f (= g)$
5. Else if $f, g \in A$,
   if $\langle f, g \rangle \in C_A$, then return $\langle f, g \rangle$
   else return $TOP$
   end.
6. Else if $f = l : a_f$ and $g = l : a_g$
   and $l \in L \cup C_L$,
   if ( $a_{fg} :=$ )UNIFY$(a_f, a_g)$,
   then return $l : a_{fg}$
   else return $TOP$
   end.
7. Else if $f = l_f : a_f$ and $g = l_g : a_g$
   and $\langle l_f, l_g \rangle \in C_L$
   and ( $a_{fg} :=$ )UNIFY$(a_f, a_g)$,
   then return $\langle l_f, l_g \rangle : a_{fg}$
8. Else if $f = f_1 \wedge f_2$
   and ( $\ll h, f_r, g_r \gg :=$ )UNIFY-CONJ$(f, g)$
   and ( $h_r :=$ )UNIFY$(f_r, g_r)$,
   then return $h \wedge h_r$
9. Else if $g = g_1 \wedge g_2$, then return UNIFY$(g, f)$
10. Else return $f \wedge g$
    end.

Function UNIFY-CONJ$(f, g)$ returns one possible 3-tuple of *feature descriptions* $\ll h, f_r, g_r \gg$: where $f$ and $g$ are *feature descriptions*, and $h$ is a unified *feature description*, and $f_r, g_r$ are rest parts of $f, g$ that are not used to generate $h$.

1. If $f = f_1 \wedge f_2$,
   ( $\ll h, f_r, g_r \gg :=$)UNIFY-CONJ$(f_1, g)$
   and return $\ll h, f_r \wedge f_2, g_r \gg$
   or
   ( $\ll h, f_r, g_r \gg :=$)UNIFY-CONJ$(f_2, g)$
   and return $\ll h, f_1 \wedge f_r, g_r \gg$
2. Else if $g = g_1 \wedge g_2$
   and ( $\ll h, g_r, f_r \gg :=$)UNIFY-CONJ$(g, f)$
   then return $\ll h, f_r, g_r \gg$
3. Else ( $h :=$ )UNIFY$(f, g)$
   and return $\ll h, NIL, NIL \gg$
   end.

## 2.2 Unification of Feature Descriptions of Two Languages

Feature Descriptions of translation examples of both languages are generated by syntactic analysis. A translation example is given in Example 2.

**Example 2**

E: I wrote a letter with a pencil.
J: 私 (I) は (topic) 鉛筆 (pencil) で (case-marker) 手紙 (letter) を (case-marker) 書いた (wrote)。

From the English sentence of this example, two feature descriptions below are generated because of the ambiguity caused by pp-attachment.

$$\begin{bmatrix} pred : write \\ tense : past \\ subj : \begin{bmatrix} pred : I \end{bmatrix} \\ obj : \begin{bmatrix} pred : letter \\ spec : a \end{bmatrix} \\ with : \begin{bmatrix} pred : pencil \\ spec : a \end{bmatrix} \end{bmatrix}$$

$$\begin{bmatrix} pred : write \\ tense : past \\ subj : \begin{bmatrix} pred : I \end{bmatrix} \\ obj : \begin{bmatrix} pred : letter \\ spec : a \\ with : \begin{bmatrix} pred : pencil \\ spec : a \end{bmatrix} \end{bmatrix} \end{bmatrix}$$

[2] These compatibility sets do not necessarily define equivalence relations of atomic values and feature labels, i.e., they do not satisfy the transitive and symmetric laws. They are reflexive, and $\langle a, a \rangle$ and $\langle l, l \rangle$ are identified as $a$ and $l$.

[3] In fact, in the case of the unification of feature descriptions of two languages, $a_i$ of $\langle a_i, a_j \rangle (\in C_A)$ is an atomic value of one language and $a_j$ is an atomic value of the other language. This is also the case with $l_i$ and $l_j$ of $\langle l_i, l_j \rangle (\in C_L)$.

From the Japanese sentence, the following single feature description is generated.

$$\begin{bmatrix} pred : 書く \\ tense : past \\ は : \begin{bmatrix} pred : 私 \end{bmatrix} \\ を : \begin{bmatrix} pred : 手紙 \end{bmatrix} \\ で : \begin{bmatrix} pred : 鉛筆 \end{bmatrix} \end{bmatrix}$$

### Set of Compatible Pairs of Atomic Values

Knowledge about word equivalence is extracted from bilingual dictionaries in order to construct $C_A$. First, for each word in the English sentence, equivalent Japanese words are extracted from English-Japanese dictionaries, and for each word in the Japanese sentence, equivalent English words are extracted from Japanese-English dictionaries[4]. Using this knowledge, any possible pairs of equivalent content words[5] that are included in the original sentences are collected, and $C_{AD}$, the set of these equivalent (i.e. compatible) word pairs, is constructed. Then for all other content words $W_{NDeng}$ in the English sentence and $W_{NDJap}$ in the Japanese sentence, any possible pairs $\langle W_{NDeng}, W_{NDjap} \rangle$ are collected, which comprise $C_{AND}$. Finally, $C_A$ is defined as $C_{AD} \cup C_{AND}$.

In the case of Example 2, $C_{AD}$, $C_{AND}$ and $C_A$ are shown below. $C_{AD}$ and $C_{AND}$ are constructed only for the content words, so in this case $C_{AND}$ is $\emptyset$ (an empty set).

$$C_{AD} = \{\langle write, 書く \rangle, \langle I, 私 \rangle, \langle letter, 手紙 \rangle, \langle pencil, 鉛筆 \rangle\},$$

$$C_{AND} = \emptyset, \quad C_A = C_{AD} \cup C_{AND}$$

### Set of Compatible Pairs of Feature Labels

In our framework of unification between two languages, we assume that the set of compatible pairs of feature labels, $C_L$, is constructed based on statistical data. That is, each feature label pair $\langle l_i, l_j \rangle$ in $C_L$ has a probability $p_{ij} (0 < p_{ij} \leq 1)$ calculated from statistical data. This $p_{ij}$ represents the probability that the semantic role of feature $l_i$ in a specific feature description of one language is the same as that of feature $l_j$ in another specific feature description of the other language. For example, for a specific English-Japanese verb pair $\langle write, 書く \rangle$, the feature label pair $\langle subj, が \rangle$ is assumed to have a probability $p_{\langle subj, が \rangle}$. And for another English-Japanese verb pair $\langle read, 読む \rangle$, the feature label pair $\langle subj, が \rangle$ is assumed to have another probability $q_{\langle subj, が \rangle}$.

Since we are at the starting point of our project of lexical knowledge acquisition, we initially assign 1 to the probability of each feature label pair, except for pairs that are known not to have the same case role from some grammatical knowledge. These exceptional pairs are not contained in $C_L$, i.e., their probabilities are 0. In fact, for the purpose of lexical knowledge acquisition, it is sufficient to assume the probability as 1 or 0, because we need credible results for extracting lexical knowledge about the usages of words.

### The Most Overlapping Unifier

The scoring function SCORE($h$) calculates the validity of a unified feature description $h$. This function returns a 2-tuple of real numbers[6], $\langle x_1, x_2 \rangle$ $(x_1, x_2 \in R(\textit{set of real numbers}))$, where $x_1$ is the number of word pairs extracted from bilingual dictionaries and contained in the unified feature description, on the other hand $x_2$ is the number of word pairs also contained in the unified feature description but not extracted from bilingual dictionaries. More precisely, $x_1$ corresponds to the number of word pairs $\langle W_{Deng}, W_{Djap} \rangle$ in the unified feature description that are elements of $C_{AD}$, and $x_2$ corresponds to the number of word pairs $\langle W_{NDeng}, W_{NDjap} \rangle$ in the unified feature description that are elements of $C_{AND}$.

The order among scores is defined as follows:

$\langle x_1, x_2 \rangle$ is greater than $\langle y_1, y_2 \rangle$

iff. $x_1 > y_1$ or $(x_1 = y_1, x_2 > y_2)$

The most overlapping unifiers are the ones with the greatest score. The complete definition of the scoring function is given below.

Function SCORE($h$) returns $\langle x_1, x_2 \rangle$ $(x_1, x_2 \in R(\textit{set of real numbers}))$:
where $h$ is a unified *feature description*.

1. If $h \in C_{AD}$, then return $\langle 1, 0 \rangle$
2. Else if $h \in C_{AND}$, then return $\langle 0, 1 \rangle$
3. Else if $h = l : a$ where $l \in L \cup C_L$ and $a \in A \cup C_A$ and SCORE($a$) = $\langle x_1, x_2 \rangle$, then return $\langle \text{SCORE}_L(l) \times x_1, \text{SCORE}_L(l) \times x_2 \rangle$
4. Else if $h = h_1 \wedge h_2$ where $h_1, h_2 \in$ FDLC and SCORE($h_1$)= $\langle x_{11}, x_{12} \rangle$ and SCORE($h_2$)= $\langle x_{21}, x_{22} \rangle$, then return $\langle x_{11} + x_{21}, x_{12} + x_{22} \rangle$
5. Else return $\langle 0, 0 \rangle$

end.

Function SCORE$_L(l)$ returns the probability of $l$: where $l \in L \cup C_L$

1. If $l \in L$, then return 1
2. If $l \in C_L$, then return the probability of $l$

[4] At present, we use a Japanese-English dictionary only, which has about 50,000 entries.

[5] Words are divided into two categories: content words and functional words. Content words are ones which can be the head of a phrase, such as nouns and verbs.

[6] Since the probability of a feature label pair is 1 or 0, $x_1$ and $x_2$ are integers at present.

Example

The results of unification and scoring of Example 2 are as below.

$$score = \langle 4, 0 \rangle$$
$$\left[\begin{array}{l} pred : \langle write, 書く \rangle \\ tense : past \\ \langle subj, は \rangle : \left[ pred : \langle I, 私 \rangle \right] \\ \langle obj, を \rangle : \left[\begin{array}{l} pred : \langle letter, 手紙 \rangle \\ spec : a \end{array}\right] \\ \langle with, で \rangle : \left[\begin{array}{l} pred : \langle pencil, 鉛筆 \rangle \\ spec : a \end{array}\right] \end{array}\right]$$

$$score = \langle 3, 0 \rangle$$
$$\left[\begin{array}{l} pred : \langle write, 書く \rangle \\ tense : past \\ \langle subj, は \rangle : \left[ pred : \langle I, 私 \rangle \right] \\ \langle obj, を \rangle : \left[\begin{array}{l} pred : \langle letter, 手紙 \rangle \\ spec : a \\ with : \left[\begin{array}{l} pred : pencil \\ spec : a \end{array}\right] \end{array}\right] \\ で : \left[ pred : 鉛筆 \right] \end{array}\right]$$

The prepositional phrase *"with a pencil"* modifies the verb *"wrote"* in the upper feature description. The score of the upper feature description is greater than that of the lower one. So in this case, the upper one is regarded as the correct case frame example for the pair $\langle write, 書く \rangle$.

# 3 Syntactic Disambiguation: Experiment and Evaluation

In order to evaluate how well syntactic ambiguities of translation examples are resolved, we made an experiment of syntactic disambiguation using 189 translation examples extracted from a Japanese-English dictionary. Firstly, each sentence of a translation example is syntactically analyzed and translated into feature descriptions. For 44 translation examples, syntactic analysis of the Japanese or English sentence is failed. For those which are successfully analyzed, the average number of feature descriptions generated from one sentence is 4.4 for Japanese and 17.1 for English. Secondly, these feature descriptions are unified. After this process of syntactic disambiguation, from 86 translation examples, a unique case frame of the unified verb pair of Japanese and English is acquired. Calculating from this result, the success ratio of acquiring unified case frames of verbs, (the number of translation examples such that a unique unified case frame of verbs is acquired from each translation example)/ (the number of translation examples such that each sentence is successfully analyzed), is $86/145 = 59.3\%$. And the success ratio of syntactic disambiguation, (the number of sentences such that a unique case frame of the verb is acquired from more than one feature descriptions)/ (the number of sentences such that more than one feature descriptions are originally generated), is $70/103 = 68.0\%$ for Japanese, and $84/133 = 63.2\%$ for English.

# 4 Lexical Knowledge Acquisition of Verbs

## 4.1 Acquiring Case Frames of Verbs

As described in 2.2, a feature description unified between English and Japanese is as below.

$$\left[\begin{array}{l} pred : \langle write, 書く \rangle \\ tense : past \\ \langle subj, は \rangle : \left[ pred : \langle I, 私 \rangle \right] \\ \langle obj, を \rangle : \left[\begin{array}{l} pred : \langle letter, 手紙 \rangle \\ spec : a \end{array}\right] \\ \langle with, で \rangle : \left[\begin{array}{l} pred : \langle pencil, 鉛筆 \rangle \\ spec : a \end{array}\right] \end{array}\right]$$

This feature description tells that the verbal concept represented by the pair of the English verb *"write"* and the Japanese verb "書く" have at least three cases that are marked by some syntactic information and some surface functional words such as $\langle subj, は \rangle$, $\langle obj, を \rangle$, $\langle with, で \rangle$. It also tells that each case takes a certain nominal concept represented by the pair of English and Japanese words, such as $\langle I, 私 \rangle$, $\langle letter, 手紙 \rangle$, $\langle pencil, 鉛筆 \rangle$. Once a large amount of this kind of data is collected, statistical data about case frames of verbs can be extracted, making use of a thesaurus of nominal concepts[7]. In the remainder of this section, we will illustrate a general procedure for acquiring case frames of verbs.

Let us start with a collection of a large amount of unified feature descriptions like above for a specific Japanese verb $V_J$. Suppose that we want to get possible case frames of this verb. By a case frame, we mean something like a feature description for this verb, consisting of surface cases each of which is marked by a postpositional particle $p_J$ and some specific semantic categories taken from a thesaurus like BGII. Usually, a verb has several distinct case frames. However, it is not easy to extract those case frames automatically only from the collected unified feature descriptions. So the system finds critical points to distinguish possible case frames for a verb using some heuristics, then it asks the human instructor whether the distinctions of case frames are correct. These heuristics and human interactions are summarized as follows.

[7] At present, an on-line thesaurus called 'Bunrui Goi Hyou'(BGH)[8] is available for Japanese. BGH has a six-layered abstraction hierarchy and more than 60,000 words are assigned at the leaves. At the present stage, it is not certain whether this thesaurus is reliable enough for our initial research target of acquiring case frames of verbs. It is, however, the most precise and broad covering Japanese thesaurus obtainable for us, currently.

**Heuristics**

1. *Semantic Category in a Thesaurus*

   First, collect the nouns marked by $p_J$ in a feature description of the verb $V_J$ from the set of unified feature descriptions. Then mark each collected noun in the thesaurus. If the most specific common layer of the marked nouns is low enough, then we assume that the case marked by $p_J$ takes a noun of the semantic category that corresponds to that layer. But if the most specific common layer is higher than a predetermined layer[8], the information provided by that layer is too general for the semantic categories of the case marked by $p_J$. For instance, it is quite rare that both an animate concept and an abstract concept can be the subject of a certain verb. Such a case strongly suggests that the verb has at least two distinct conceptual meanings or two distinct case frames. It then becomes necessary to classify the marked nouns in the thesaurus.

2. *Bilingual Intersection of Concepts*

   Some of the heuristics come from the advantages of bilingual intersection of concepts, which we have already shown in Chapter 1 as *semantic disambiguation*. For a Japanese verb $V_J$ and its case marked by a postpositional particle $p_J$, suppose that unified feature descriptions such as [ $pred{:}\langle V_{E1}, V_J\rangle$, $\langle l_{E1}, p_J\rangle{:}\langle N_{E1}, N_{J1}\rangle$ ] and [ $pred{:}\langle V_{E2}, V_J\rangle$, $\langle l_{E2}, p_J\rangle{:}\langle N_{E2}, N_{J2}\rangle$ ] are obtained. Both of these two feature descriptions have a feature label $p_J$ for $V_J$. However, if $V_{E1}$ and $V_{E2}$ are different verbs or $l_{E1}$ and $l_{E2}$ are different feature labels, these two feature descriptions may be classified into different case frames of the verb $V_J$.

3. *Correlation of Cases*

   Another heuristics are related to sentence patterns of verbs. Sometimes the case marked by $p_J$ has a correlation with other cases in sentence patterns. If the correlations between cases are detected, then it helps the classification, and some sentence patterns (or case frames) of the verb $V_J$ will be acquired.

**Human Interactions**

As described above, the system can find critical points to distinguish possible case frames for a verb by those heuristics. The system, however, cannot determine the distinction only with positive data collected from examples. The main purpose of human interaction is to obtain negative examples. The system asks the human instructor whether a case marked by $p_{J1}$ and another case marked by $p_{J2}$ can co-occur or not. If they cannot co-occur, then the system learns that $V_J$ has at least two sentence patterns (or case frames) and that one of them has the case marked by $p_{J1}$ and the other has the case marked by $p_{J2}$. An example of human interactions of this type is shown in next section.

It is often said that hand-made semantic dictionary contains quite unstable data, which means that it strongly depends on the human composer. In order to acquire stable lexical knowledge base, we decided to limit human interactions to yes-no type of questions and answers, such that the system asks the human instructor whether something is true or false so that he can answer only yes or no.

[8]The predetermined layers depend on the thesaurus we are dealing with.

Table 1: Semantic Marker of IPAL

| CON | concrete | ABS | abstract |
|---|---|---|---|
| ANI | animal | ACT | action |
| HUM | human | MEN | mental |
| ORG | organization | LIN | linguistic products |
| PLA | plant | CHA | character |
| PAR | parts | REL | relation |
| NAT | natural | LOC | location |
| PRO | products | TIM | time |
| | | QUA | quantity |
| PHE | phenomenon | DIV | diverse |

Table 2: Acquired Case Slots for "書く *(write)*"

| Case Slots | Sem. Mark. | Freq. | Examples |
|---|---|---|---|
| ⟨*subj*, は・が⟩ | HUM | 95 | 私 (I) |
| ⟨*obj*, は・を⟩<br>⟨[*subj*, *passive*], は・が⟩ | REL,<br>QUA,<br>LIN | 153 | 手紙 (letter)、<br>名前 (name) |
| ⟨*with*, で⟩ | PRO | 10 | ペン (pen) |
| ⟨*in*, で⟩ | LIN,<br>REL | 28 | 漢字 (kanji)<br>形式 (form) |
| ⟨*on*, に⟩ | PRO | 16 | 紙 (paper) |
| ⟨*to*, に⟩ | HUM | 13 | 父 (father) |

## 4.2 Examples and Evaluations

We have collected about 50,000 translation examples from a machine readable Japanese-English dictionary and an English learners' textbook. In this bilingual corpus, about 70 distinct Japanese verbs appear in more than 100 examples. We have obtained unified feature descriptions for several verbs which appeared more than 200 times. From them we have gotten some case frames. In this experiment we used the set of semantic markers defined in IPAL [4], listed in Table 1.

Table 2 shows the case slots of "書く *(write)*" extracted from 207 translation examples. In the process of extraction, bilingual feature label pairs are quite useful to find different case slots that are marked by the same postpositional particle in Japanese. In order to acquire case frames of the verb "書く *(write)*" from

Table 3: Acquired Case Frames for "書く *(write)*"

| Case Frame 1 | | Case Frame 2 | |
|---|---|---|---|
| に *(on)* | PRO | に *(to)* | HUM |
| は・が *(subj)* | HUM | は・が *(subj)* | HUM |
| は・を *(obj)* は・が ( *[subj, passive]*) | REL, QUA, LIN | は・を *(obj)* は・が ( *[subj, passive]*) | LIN |
| で *(with)* | PRO | で *(with)* | PRO |
| で *(in)* | LIN, REL | で *(in)* | LIN, REL |

the extracted case slots, the system asks the human instructor about the possibilities of the co-occurrence of the case slots that do not co-occur in the translation examples by composing sample phrases. The questions and answers are as follows.

**QUESTION 1 :**
*Can I say*
*"ペン (pen) で (with) 英語 (English) で (in) 書く (write)" ?*
⟶ ⟶ ⟶ *YES.*

**QUESTION 2 :**
*Can I say*
*"カード (card) に (on) 父 (father) に (to) 書く (write)" ?*
⟶ ⟶ ⟶ *NO.*

The postpositional particle "に" is used to mark two different cases of the verb "書く *(write)*" in Japanese sentences. One of them represents things on which something is written like in *"write something on a sheet of paper"*, and the other represents someone to whom a correspondence is written, like in *"write a letter to a lover"*. The difference of these two usages is clear by the bilingual feature label pairs ⟨*on*, に⟩ and ⟨*to*, に⟩. The human instructor answers that only these two case slots cannot co-occur. Then two case frames are obtained as in Table 3.

This simple experiment suggests that it is quite possible to acquire case frames of verbs from bilingual corpora if enough translation examples are available. Actually, on the assumption that 200 translation examples are necessary for acquiring case frames of one verb, 100,000 translation examples are necessary for 70 verbs. If a bilingual corpus of 1,000,000 translation examples is obtained, it is possible to compile a semantic dictionary with the same scale as IPAL through a little interaction with a human instructor for each verb. We think it possible to construct a bilingual corpus of that scale or more in the near future.

# 5 Concluding Remarks

We have proposed a method for resolving the syntactic ambiguities of translation examples of bilingual corpora and a method for acquiring case frames of verbs. At present, we are extending our prototype system for acquiring case frames of verbs, and the detail of the extended system will be reported in the future. We believe that the proposed method is applicable to several other problems as well. One of them is to acquire features of nominal concepts. We are at the moment looking at some specific nominal expression *"A の B"* in Japanese, corresponding literally to *"B of A"* in English. That expression specifies a variety of relationships of noun phrases, which are often stated in different expressions in English. They will help to acquire typical attributes of nominal concepts from bilingual corpora. Our method is also useful to collect parsed translation examples for example-based translation [9] and to acquire translation patterns between two languages.

# References


[1] Brent, M.: "Automatic Acquisition of Subcategorization Frames from Untagged Text", *Proc. of the 29th Annual Meeting of the ACL*, 1991.

[2] Dagan, I., Itai, A. and Schwall, U.: "Two Languages are More Informative Than One", *Proc. of the 29th Annual Meeting of the ACL*, 1991.

[3] Hindle, D.: "Noun Classification from Predicate Argument Structures", *Proc. of the 28th Annual Meeting of the ACL*, 1990.

[4] Information-technology Promotion Agency, Japan: *IPA Lexicon of the Japanese Language for computers IPAL (Basic Verbs)*,(in Japanese), 1987.

[5] Japan Electric Dictionary Research Institute, Ltd.: *Conceptual Dictionary, 2nd. Edition*,(in Japanese), TR-012, 1989.

[6] Kasper, R. and Rounds, W.: "A Logical Semantics for Feature Structures", *Proc. of the 24th Annual Meeting of the ACL*, 1986.

[7] Lenat, D. et al.: *Building Large Knowledge-based Systems*, Addison-Wesley, 1990.

[8] National Language Research Institute: *Word List by Semantic Principles*,(in Japanese), Syuei Syuppan, 1964.

[9] Sato, S. and Nagao, M.: "Memory-based Translation", *Proc. of the 13th International Conference on Computational Linguistics*, Helsinki, Finland, 1990.

[10] Shimizu, M. and Narita, N., eds.: *Japanese-English Dictionary*, Kodansha Gakujutsu Bunko, 1979.

[11] Tokunaga, T. and Tanaka, H.: "The Automatic Extraction of Conceptual Items from Bilingual Dictionaries", (in Japanese), *Journal of Japan Society for Artificial Intelligence*, Vol.6, No.2, 1991.

[12] Tomiura, Y., Hitaka, T. and Yoshida, S.: "Extracting the Superordinate-Subordinate Relation between Verbs from Definition Sentences in Japanese Dictionaries", (in Japanese), *Journal of Information Processing, IPSJ*, Vol.32, No.1, 1991.

[13] Tsurumaru, H., Takesita, K., Itami, K., Yanagawa, T. and Yoshida, S.: "An Approach to Thesaurus Construction",(in Japanese), *IPSJ-WGNL*, 83-16, 1991.