Sense Classification of Verbal Polysemy based-on Bilingual Class/Class Association*

Takehito Utsuro
Graduate School of Information Science, Nara Institute of Science and Technology
8916-5, Takayama-cho, Ikoma-shi, Nara, 630-01, JAPAN
utsuro@is.aist-nara.ac.jp

Abstract

In the field of statistical analysis of natural language data, the measure of word/class association has proved to be quite useful for discovering a meaningful sense cluster in an arbitrary level of the thesaurus. In this paper, we apply its idea to the sense classification of Japanese verbal polysemy in case frame acquisition from Japanese-English parallel corpora. Measures of *bilingual class/class association* and *bilingual class/frame association* are introduced and used for discovering sense clusters in the sense distribution of English predicates and Japanese case element nouns. In a small experiment, 93.3% of the discovered clusters are *correct* in that none of them contains examples of more than one hand-classified senses.

1 Introduction

In corpus-based NLP, acquisition of lexical knowledge has become one of the major research topics. Among several research topics in this field, acquisition from parallel corpora is quite attractive (e.g. Dagan et al. (1991)). The reason is that parallel sentences are useful for resolving both syntactic and lexical ambiguities in the monolingual sentences. Especially if the two languages have different syntactic structures and word meanings (such as English and Japanese), this approach has proved to be most effective in disambiguation (Matsumoto et al., 1993; Utsuro et al., 1993).

Utsuro et al. (1993) proposed a method for acquiring surface case frames of Japanese verbs from Japanese-English parallel corpora. In this method, translated English verbs and case labels are used to classify senses of Japanese polysemous verbs. Clues to sense classification are found using English verbs and case labels, as well as the sense distribution of the Japanese case element nouns. Then, a human instructor judges whether the clues are correct. One of the major disadvantages of the method is that the use of English information and sense distribution of Japanese case element nouns is restricted. Only surface forms of English verbs and case labels are used and sense distribution of English verbs is not used. Also, the threshold of deciding a distinction in the sense distribution of Japanese case element nouns is predetermined on a fixed level in a Japanese thesaurus. As a result, the human instructor is frequently asked to judge the correctness of the clue.

In the field of statistical analysis of natural language data, it is common to use measures of lexical association, such as the information-theoretic measure of mutual information, to extract useful relationships between words (e.g. Church and Hanks (1990)). Lexical association has its limits, however, since often either the data is insufficient to provide reliable word/word correspondences, or the task requires more abstraction than word/word correspondences permit. Thus, Resnik (1992) proposed a useful measure of word/class association by generalizing information-theoretic measure of word/word association. The proposed measure addresses the limitations of lexical association by facilitating statistical discovery of facts involving word *classes* rather than individual words.

We find the measure of word/class association of Resnik (1992) is quite attractive, since it is possible to discover a meaningful sense cluster in an arbitrary level of the thesaurus. We thus expect that the restrictions of the previous method of Utsuro et al. (1993) can be overcome by employing the idea of the measure of word/class association. In this paper, we describe how this idea can be applied to the sense classification of Japanese verbal polysemy in case frame acquisition from Japanese-English parallel corpora. First, sense distribution of English predicates and Japanese case element nouns is represented using monolingual English and Japanese thesaurus, respectively (sections 2 and 3). Then, the measure of the association of classes of English predicates and Japanese case element nouns, i.e., a measure of *bilingual class/class association*, is introduced, and extended into a measure of *bilingual class/frame association* (section 4).

*The author would like to thank Prof. Yuji MATSUMOTO for his valuable comments on this research. This work is partly supported by the Grants from the Ministry of Education, Science, and Culture, Japan, #07780326.

Using these measures, sense clusters are discovered in the sense distribution of English predicates and Japanese case element nouns. Finally, examples of a Japanese polysemous verb collected from Japanese-English parallel corpora are divided into disjoint clusters according to those discovered sense clusters (section 5). The results of a small experiment are presented and the proposed measure is evaluated (section 6).

2 Bilingual Surface Case Structure

In the framework of verbal case frame acquisition from parallel corpora, bilingually matched surface case structures (Matsumoto et al., 1993) are collected and surface case frames of Japanese verbs are acquired from the collection. In this paper, each bilingually matched surface case structure is called a *bilingual surface case structure*, and represented as a feature structure:

$$\left[\begin{array}{l} pred : v_J \\ sem_E : SEM_E \\ p_1 : \left[\begin{array}{l} pred : n_{J1} \\ sem : SEM_{J1} \end{array}\right] \\ \quad \vdots \qquad \vdots \\ p_m : \left[\begin{array}{l} pred : n_{Jm} \\ sem : SEM_{Jm} \end{array}\right] \end{array}\right]$$

v_J indicates the verb in the Japanese sentence, $p_1, \ldots, p_m$ denote the Japanese case markers, and $n_{J1}, \ldots, n_{Jm}$ denote the Japanese case element nouns. When a Japanese noun n_{Ji} has several senses, it may appear in several leaf classes in the Japanese thesaurus. Thus, SEM_{Ji} is represented as a set of those classes, and is referred to as a *semantic label*. SEM_E is a semantic label of the corresponding English predicate, i.e., a set of classes in the English thesaurus:

$$SEM_E = \{c_{E1}, \ldots, c_{Ek}\},\ SEM_{Ji} = \{c_{J1}, \ldots, c_{Jl}\}$$

$c_{E1}, \ldots, c_{Ek}$ and $c_{J1}, \ldots, c_{Jl}$ indicate the classes in the English and Japanese thesaurus, respectively.

By structurally matching the Japanese-English parallel sentences in Example 1, the following bilingual surface case structure is obtained:

Example 1

J:	Watashi-ha	uwagi-wo	kagi-ni	kaketa.	
	I-TOP	*coat-ACC*	*hook-on*	*hung*	
E:	I hung my coat on the hook.				

$$\left[\begin{array}{l} pred : kakeru \\ sem_E : \{c_{hang1}, \ldots, c_{hang4}\} \\ ha : \left[\begin{array}{l} pred : watashi \\ sem : \{c_w\} \end{array}\right] \\ wo : \left[\begin{array}{l} pred : uwagi \\ sem : \{c_u\} \end{array}\right] \\ ni : \left[\begin{array}{l} pred : kagi \\ sem : \{c_{k1}, \ldots, c_{k4}\} \end{array}\right] \end{array}\right]$$

We use Roget's Thesaurus (Roget, 1911) as the English thesaurus and 'Bunrui Goi Hyou'(BGH) (NLRI, 1993) as the Japanese thesaurus. In Roget's Thesaurus, the verb *"hang"* has four senses. In BGH, the nouns *"watashi"* and *"uwagi"* have only one sense, respectively, and *"kagi"* has four senses.

3 Monolingual Thesaurus

A thesaurus is regarded as a tree in which each node represents a class. We introduce $\preceq$ as the superordinate-subordinate relation of classes. In general, $c_1 \preceq c_2$ means that c_1 is subordinate to c_2. We define $\preceq$ so that a semantic label $SEM = \{c_1, \ldots, c_n\}$ is subordinate to each class c_i:

$$\forall c \in SEM, \quad SEM \preceq c$$

When searching for classes which give maximum association score (section 5), this definition makes it possible to calculate association score for all the senses in a semantic label and to find senses which give a maximum association score[1].

BGH has a six-layered abstraction hierarchy and more than 60,000 Japanese words are assigned at the leaves and its nominal part contains about 45,000 words[2]. Roget's Thesaurus has a seven-layered abstraction hierarchy and over 100,000 words are allocated at the leaves[3]. In Roget's Thesaurus, sense classification is preferred to part of speech distinction. Thus, a noun and a verb which have similar senses are assigned similar classes in the thesaurus.

4 Class-based Association Score

4.1 Word/Class Association Score

The measure of word/class association of Resnik (1992) can be illustrated by the problem of finding the prototypical object classes for verbs. Let $\mathcal{V}$ and $\mathcal{N}$ be the sets of all verbs and nouns, respectively. Given a verb $v(\in \mathcal{V})$ and a noun class $c(\subseteq \mathcal{N})$, the joint probability of v and c is estimated as

$$\Pr(v, c) \approx \frac{\sum_{n \in c} count(v, n)}{\sum_{v' \in \mathcal{V}} \sum_{n' \in \mathcal{N}} count(v', n')}$$

The *association score* $A(v, c)$ of a verb v and a noun class c is defined as

$$A(v, c) = \Pr(c \mid v) \log \frac{\Pr(v, c)}{\Pr(v)\Pr(c)} = \Pr(c \mid v)\, \mathrm{I}(v; c)$$

The association score takes the mutual information between the verb and a noun class, and scales

[1] This process corresponds to sense disambiguation by maximizing the association score.

[2] Five classes are allocated at the next level from the root node: *abstract-relations, agents-of-human-activities, human-activities, products,* and *natural-objects-and-natural-phenomena.*

[3] At the next level from the root node, it has six classes: *abstract-relations, space, matter, intellect, volition,* and *affections.*

it according to the likelihood that a member of the class will actually appear as the object of the verb. The first term of the conditional probability measures the generality of the association, while the second term of the mutual information measures the co-occurrence of the association.

4.2 Bilingual Class/Class Association Score

We now apply the word/class association score to the task of measuring the association of classes of English predicates and Japanese case element nouns in the collection of bilingual surface case structures. First, we assume that for any polysemous Japanese verb v_J, there exists a case marker p which is most effective for sense classification of v_J. Given the collection of bilingual surface case structures for v_J, we introduce the *bilingual class/class association score* for measuring the association of a class c_E of English predicates and a class c_J of Japanese case element nouns for a case marker p.

Let $Eg(v_J,p)$ be the set of bilingual surface case structures collected from the Japanese-English parallel corpora, each element of which has a Japanese verb v_J and a Japanese case marker p. Among the elements of $Eg(v_J,p)$, let $Eg(v_J,p,c_E)$ be the set of those whose semantic label SEM_E of the English predicate satisfies the class c_E, i.e., $SEM_E \preceq c_E$, and $Eg(v_J,p/c_J)$ be the set of those whose semantic label SEM_J of the Japanese case element noun for the case marker p satisfies the class c_J, i.e., $SEM_J \preceq c_J$. Let $Eg(v_J,c_E,p/c_J)$ be the intersection of $Eg(v_J,p,c_E)$ and $Eg(v_J,p/c_J)$. Then, conditional probabilities $\Pr(c_E \mid v_J,p)$, $\Pr(c_J \mid v_J,p)$, and $\Pr(c_E,c_J \mid v_J,p)$ are defined as the ratios of the numbers of the elements of those sets:

$$\Pr(c_E \mid v_J,p) = \frac{|Eg(v_J,p,c_E)|}{|Eg(v_J,p)|}$$

$$\Pr(c_J \mid v_J,p) = \frac{|Eg(v_J,p/c_J)|}{|Eg(v_J,p)|}$$

$$\Pr(c_E,c_J \mid v_J,p) = \frac{|Eg(v_J,c_E,p/c_J)|}{|Eg(v_J,p)|}$$

Then, given v_J and p, the association score $A(c_E,c_J \mid v_J,p)$ of c_E and c_J is defined as

$$A(c_E,c_J \mid v_J,p) = \Pr(c_E,c_J \mid v_J,p) \log \frac{\Pr(c_E,c_J \mid v_J,p)}{\Pr(c_E \mid v_J,p)\Pr(c_J \mid v_J,p)}$$

This definition is slightly different from that of the word/class association score in that it only needs the set $Eg(v_J,p)$ for a Japanese verb v_J and a Japanese case marker p, but not the whole Japanese-English parallel corpora. This is because our task is to discover strong association of an English class and a Japanese class in $Eg(v_J,p)$, rather than in the whole Japanese-English parallel corpora. Besides, as the first term for measuring the generality of the association, we use $\Pr(c_E,c_J \mid v_J,p)$ instead of $\Pr(c_J \mid v_J,p,c_E)$ or $\Pr(c_E \mid v_J,p/c_J)$ below:[4]

$$\Pr(c_J \mid v_J,p,c_E) = \frac{|Eg(v_J,c_E,p/c_J)|}{|Eg(v_J,p,c_E)|}$$

$$\Pr(c_E \mid v_J,p/c_J) = \frac{|Eg(v_J,c_E,p/c_J)|}{|Eg(v_J,p/c_J)|}$$

4.3 Bilingual Class/Frame Association Score

In the previous section, we assume that for any polysemous Japanese verb v_J, there exists a case marker p which is most effective for sense classification of verbal polysemy v_J. However, it can happen that a combination of more than one case marker characterizes a sense of the verbal polysemy v_J. Even if there exists exactly one case marker which is most effective for sense classification, it is necessary to select the most effective case marker automatically by some measure. For example, using some measure, it is desirable to automatically discover the fact that, for the task of sense classification of verbal polysemy, subject nouns are usually most effective for intransitive verbs, while object nouns are usually most effective for transitive verbs.

This section generalizes the previous definition of *bilingual class/class association score*, and introduces the *bilingual class/frame association score*. In the new definition, we consider every possible set of pairs of a Japanese case marker p and a Japanese noun class c_J, instead of predetermining the most effective case marker. The *bilingual class/frame association score* measures the association of an English class c_E and a set of pairs of a Japanese case marker p and a Japanese noun class c_J marked by p. By searching for a large association score, it becomes possible to find any combination of case markers which characterizes a sense of the verbal polysemy v_J.

4.3.1 Japanese Case-Class Frame

First, we introduce a data structure which represents a set of pairs of Japanese case marker p and a Japanese noun class c_J marked by p, and call it *Japanese case-class frame*. A *Japanese case-class frame* can be represented as a feature structure:

$$\begin{bmatrix} p_1 : c_{J1} \\ \vdots \\ p_m : c_{Jm} \end{bmatrix}$$

[4] $\Pr(c_J \mid v_J,p,c_E)$ and $\Pr(c_E \mid v_J,p/c_J)$ are too large in lower parts of the thesaurus, since we focus on examples which have a Japanese verb v_J and a Japanese case marker p. When we used the average of $\Pr(c_J \mid v_J,p,c_E)$ and $\Pr(c_E \mid v_J,p/c_J)$ instead of $\Pr(c_E,c_J \mid v_J,p)$ in the experiment of section 6, most discovered clusters consisted of only one example.

4.3.2 Subsumption Relation

Next, we introduce *subsumption relation* $\preceq_f$ of a *bilingual surface case structure* e and a *Japanese case-class frame* f_J:

$e \preceq_f f_J$ iff. for each case marker p in f_J and its noun class c_J, there exists the same case marker p in e and its semantic label SEM_J is subordinate to c_J, i.e. $SEM_J \preceq c_J$

This definition can be easily extended into a subsumption relation of Japanese case-class frames.

4.3.3 Bilingual Class/Frame Association Score

Let $Eg(v_J)$ be the set of bilingual surface case structures collected from the Japanese-English parallel corpora, each element of which has a Japanese verb v_J. Among the elements e of $Eg(v_J)$, let $Eg(v_J, c_E)$ be the set of those whose semantic label SEM_E of the English predicate satisfies the class c_E, i.e., $SEM_E \preceq c_E$, and $Eg(v_J, f_J)$ be the set of those which satisfy the Japanese case-class frame f_J, i.e., $e \preceq_f f_J$. Let $Eg(v_J, c_E, f_J)$ be the intersection of $Eg(v_J, c_E)$ and $Eg(v_J, f_J)$. Then, conditional probabilities $\Pr(c_E \mid v_J)$, $\Pr(f_J \mid v_J)$, and $\Pr(c_E, f_J \mid v_J)$ are defined as the ratios of the numbers of the elements of those sets:

$$\Pr(c_E \mid v_J) = \frac{|Eg(v_J, c_E)|}{|Eg(v_J)|}$$

$$\Pr(f_J \mid v_J) = \frac{|Eg(v_J, f_J)|}{|Eg(v_J)|}$$

$$\Pr(c_E, f_J \mid v_J) = \frac{|Eg(v_J, c_E, f_J)|}{|Eg(v_J)|}$$

Then, given v_J, the association score $A(c_E, f_J \mid v_J)$ of c_E and f_J is defined as

$$A(c_E, f_J \mid v_J) = \Pr(c_E, f_J \mid v_J) \log \frac{\Pr(c_E, f_J \mid v_J)}{\Pr(c_E \mid v_J)\Pr(f_J \mid v_J)}$$

As well as the case of the bilingual class/class association score, this definition only needs the set $Eg(v_J)$ for a Japanese verb v_J, not the whole Japanese-English parallel corpora.

5 Sense Classification of Verbal Polysemy

This section explains how to classify the elements of the set $Eg(v_J)$ of bilingual surface case structures according to the sense of the verbal polysemy v_J, with the bilingual class/frame association score defined in the previous section. In this classification process, pairs of an English class c_E and a Japanese case-class frame f_J which give large association score $A(c_E, f_J \mid v_J)$ are searched for. It is desirable that the set $Eg(v_J)$ be divided into disjoint subsets by the discovered pairs of c_E and f_J. The classification process proceeds according to the following steps:

1. First, the index i and the set of examples Eg are initialized as $i \leftarrow 1$ and $Eg \leftarrow Eg(v_J)$.
2. For the i-th iteration, let c_E and f_J be a pair of an English class and a Japanese case-class frame which satisfy the following constraint for all the pairs of c_{Ej} and f_{Jj} $(1 \leq j \leq i-1)$: c_E is not subordinate nor superordinate to c_{Ej} (i.e., $c_E \not\preceq c_{Ej}$ and $c_{Ej} \not\preceq c_E$), or f_J is not subordinate nor superordinate to f_{Jj} (i.e., $f_J \not\preceq_f f_{Jj}$ and $f_{Jj} \not\preceq_f f_J$). Then, among those pairs of c_E and f_J, search for a pair c_{Ei} and f_{Ji} which gives maximum association score $\max_{c_E, f_J} A(c_E, f_J \mid v_J)$,[5] and collect the elements of Eg which satisfy the restrictions of c_{Ei} and f_{Ji} into the set $Eg(v_J, c_{Ei}, f_{Ji})$.
3. Subtract the set $Eg(v_J, c_{Ei}, f_{Ji})$ from Eg as $Eg \leftarrow Eg - Eg(v_J, c_{Ei}, f_{Ji})$. If $Eg \neq \emptyset$, then increment the index i as $i \leftarrow i+1$ and go to step 2. Otherwise, set the number k of the subsets as $k \leftarrow i$ and terminate the classification process.

As the result of this classification process, the set $Eg(v_J)$ is divided into disjoint subsets $Eg(v_J, c_{E1}, f_{J1}), \ldots, Eg(v_J, c_{Ek}, f_{Jk})$.[6] For example, if a Japanese polysemous verb v_J has both intransitive and transitive senses, pairs with the *subject* case like $\langle c_{E1}, [subj : c_{J1}] \rangle, \ldots, \langle c_{Ek'}, [subj : c_{Jk'}] \rangle$ will be discovered for intransitive senses, while pairs with the *object* case like $\langle c_{Ek'+1}, [obj : c_{Jk'+1}] \rangle, \ldots, \langle c_{Ek}, [obj : c_{Jk}] \rangle$ will be discovered for transitive senses.

Given the set $Eg(v_J)$, the iterations of the association score calculation is $O(|Eg(v_J)|)$[7]. Since the classification process can be regarded as sorting the calculated association score, its computational complexity can be $O(|Eg(v_J)| \log |Eg(v_J)|)$ if efficient sorting algorithms such as quick sort are employed.

6 Experiment and Evaluation

This section gives the results of a small exper-

[5] The association score $A(c_E, f_J \mid v_J)$ is calculated from the whole set $Eg(v_J)$, not Eg.

[6] Although the classification process itself guarantees the disjointness of $Eg(v_J, c_{E1}, f_{J1}), \ldots, Eg(v_J, c_{Ek}, f_{Jk})$, the subordinate-superordinate constraint of c_E and f_J in the step 2 also guarantees the disjointness of the example sets which satisfy the restrictions of the pairs $\langle c_{E1}, f_{J1} \rangle, \ldots, \langle c_{Ek}, f_{Jk} \rangle$.

[7] Let l_J, d_J, and d_E be the maximum number of Japanese cases in a bilingual surface case structure, the depths of the Japanese and English thesauri, respectively. Then, given a bilingual surface case structure e, the number of Japanese case-class frames f_J which is superordinate to e (i.e., $e \preceq_f f_J$) is less than $2^{l_J} \times d_J^{l_J}$, and the number of possible pairs of c_E and f_J is less than $2^{l_J} \times d_J^{l_J} \times d_E$, which is constant.

Table 1: Sense Classification of *kau*

Hand-Classif.	Cluster No.	English Predicate Class (c_E)/ Japanese *wo*(ACC) Case Noun Class (c_J) (Level in the Thesaurus and Example Word)	Number of Egs.	Association Score
1	1	*buy*(Leaf)/131(Level3, *hon(book)*)	8	0.048
	2	*buy*(Leaf)/13220(Level5, *e(picture)*)	3	0.018
	3	Purchase(Leaf−1, *buy,pay*)/14(Level2, Products)	46	0.149
	4	*treat oneself to*(Leaf)/14650-6-80(Leaf, *gaisha(foreign car)*)	1	0.070
	5	*treat oneself to*(Leaf)/14280-3-10(Leaf, *yubiwa(ring)*)	1	0.070
	6	*purchase*(Leaf)/11720-3-10(Leaf,*disho(land)*)	1	0.083
	7	*bring*(Leaf)/14010-4-40(Leaf, *miyage(souvenir)*)	1	0.062
	8	*get*(Leaf)/14570-1-10(Leaf, *omocha(toy)*)	1	0.070
2	9	*incur*(Leaf)/130(Level3, *urami(enmity)*)	5	0.185
	10	Motive(Leaf−1, *rouse*)/13020-5-50(Leaf, *hankan(antipathy)*)	3	0.169
	11	*disgust*(Leaf)/13010-1-50(Leaf, *hinshuku(displeasure)*)	1	0.083
3	12	*appreciate*(Leaf)/13040-6-30(Leaf, *doryoku(effort)*)	1	0.083
	13	*get an opinion of*(Leaf)/12040-1-50(Leaf, *otoko(person)*)	1	0.083
	14	*use*(Leaf)/13421-6-50(Leaf, *shuwan(ability)*)	1	0.083
4	15	*win*(Leaf)/13010-6-200(Leaf, *kanshin(favor)*)	1	0.083
Total			75	—

Table 2: Examples of Intransitive/Transitive Distinction

Japanese Verb	English Predicate Class (c_E)/Japanese Case-Class Frame (f_J) (Level in the Thesaurus and Example Word)	Number of Egs.	Association Score
haru	*expensive*(Leaf)/*ga*(NOM):*ne(price)*(Leaf)	3	0.299
	Special Sensation(Leaf−3, *freeze*)/ *ga*(NOM):15130-11-10(Leaf, *koori(ice)*)	3	0.237
	Acts(Leaf−2, *persist,stick to*)/ *wo*(ACC):13040(Level5, *goujou(obstinacy)*)	7	0.459
hiku	Decrease(Leaf−1, *subside*)/*ga*(NOM):151(Level3, *kouzui(floods)*)	2	0.109
	Results of Reasoning(Leaf−2, *catch,have*)/ *wo*(ACC):15860-11(Level6, *kaze(cold)*)	26	0.421
hiraku	Intellect(Level1, *open*)/*ga*(NOM):14(Products)(Level2, *to(door)*)	12	0.339
	hold(Leaf)/*wo*(ACC):13510-1(Level6, *kaigou(meeting)*)	3	0.114
kanau	Completion(Leaf−1, *realize*)/*ga*(NOM):1304(Level4, *negai(desire)*)	8	0.460
	Quantity(Leaf−3, *equal*)/*ni*(DAT):12000-3-10(Leaf, *kare(he)*)	8	0.504

iment. As a Japanese-English parallel corpus, we use a corpus of about 40,000 translation examples extracted from a machine readable Japanese-English dictionary (Shimizu and Narita, 1979).

6.1 Example of *kau*

First, we show the result of classifying 75 examples (represented as bilingual surface case structures) of the Japanese polysemous verb *kau*.

As the result of searching for pairs of an English class and a Japanese case-class frame with a large association score, the *wo* case (the accusative case) is preferred as the most effective case for sense classification. 15 pairs of an English class and a Japanese case-class frame are found and the set of the 75 examples are divided into 15 disjoint clusters (Table 1). Each cluster is represented as a pair of the class c_E of the English predicates and the class c_J of the Japanese case element nouns of *wo* case, along with the level of the class in the thesaurus and the example word. English classes are taken from Roget's Thesaurus and Japanese classes from BGH[8]. In both thesauri, leaf classes correspond to one word.

For the evaluation of the results, we hand-classified the 15 clusters into four groups, each of which corresponds to only one sense of *kau*[9]. Most hand-classified clusters for *kau* consist of more than one clusters found by maximizing the association score. However, these clusters are *correct* in that none of them contains examples of more than one hand-classified senses of *kau*.

6.2 Examples of Intransitive/Transitive Distinction

For four Japanese verbs *haru, hiku, hiraku*, and *kanau*, Table 2 shows examples of classifying intransitive/transitive senses by the proposed sense

[8]The classes of BGH are represented as numerical codes, in which each digit denotes the choice of the branch in the thesaurus. The classes starting with '11', '12', '13', '14', and '15' are subordinate to *abstract-relations, agents-of-human-activities, human-activities, products* and *natural-objects-and-natural-phenomena*, respectively.

[9]The criterion of this hand-classification is taken from the existing Japanese dictionaries for human use and the hand-compiled Japanese case frame dictionary IPAL (IPA, 1987).

Table 3: Evaluation of Sense Classification

	Japanese Verb	Japanese Case-Class Frame f_J	Total Cl.	Total Eg.	One Sense Cluster Cl.	One Sense Cluster Eg.	Hand-Classif.	Total Cl. / Hand-Classif.
1	*agaru(rise)*	*ga*(NOM)	41	74	39	69 (93.2%)	17	2.41
2	*ageru(raise)*	*wo*(ACC)	54	107	52	93 (86.9%)	18	3.00
3	*aku(open,* iv*)*	*ga*(NOM)	12	29	12	29 (100%)	8	1.50
4	*haru(spread,* iv/tv*)*	*ga*(NOM)/*wo*(ACC)	19	36	17	30 (83.3%)	11	1.73
5	*hiku(subside,pull)*	*ga*(NOM)/*wo*(ACC)	40	105	40	105 (100%)	23	1.74
6	*hiraku(open,* iv/tv*)*	*ga*(NOM)/*wo*(ACC)	15	54	13	50 (92.6%)	10	1.50
7	*kakeru(hang)*	*wo*(ACC)	45	103	42	86 (83.5%)	25	1.80
8	*kanau(realize,) conform to)*	*ga*(NOM) /*ni*(DAT)	14	31	14	31 (100%)	3	4.67
9	*kau(buy)*	*wo*(ACC)	15	75	15	75 (100%)	4	3.75
Average			--	--	--	(93.3%)	--	2.46

classification method. Clusters of intransitive senses are discovered with the Japanese case-class frames which contain the *ga* case (the nominative case), while those of transitive senses are discovered with the Japanese case-class frames which contain the *wo* case (the accusative case) and *ni* case (the dative case).

6.3 Evaluation

For 9 verbs, we made an experiment on sense classification of verbal polysemy. We compared the result with the hand-classification and checked whether each cluster contained examples of only one hand-classified sense (Table 3). In the table, 'Cl.' and 'Eg.' indicate the numbers of clusters and examples, respectively. The column 'One Sense Cluster' means that each cluster contains examples of only one hand-classified sense, and the sub-columns 'Cl.' and 'Eg.' list the number of such clusters and the sum of examples contained in such clusters, respectively. We evaluated the accuracy of the method as the rate of the number of examples contained in *one sense* clusters as in the 'Eg.' sub-column. This achieved 100% accuracy for four verbs out of the 9 verbs, and 93.3% in average. The column 'Total Cl./Hand-Classif.' indicates the ratio of the total number of clusters to the number of hand-classified senses, corresponding to the average number of clusters into which one hand-classified sense is divided. Its average, median, and standard deviation are 2.46, 1.80, and 1.06, respectively.

The result of the experiment indicated that the proposed sense classification method has achieved almost pure classification, while the result seems a little finer than hand-classification. This is mainly caused by the fact that clusters which correspond to the same hand-classified sense are separately located in the human-made thesaurus, and it is not easy to find exactly one representative class in the thesaurus (Utsuro, 1995). It is necessary to further merge the clusters so that exactly one cluster corresponds to one hand-classified sense.

7 Conclusion

This paper proposed a bilingual class-based method for sense classification of verbal polysemy, which is based on the maximization of the bilingual class/frame association score. It achieved fairly high accuracy, although it is necessary to further merge the clusters so that exactly one cluster corresponds to one hand-classified sense. We are planning to make experiments on sense classification without bilingual information to evaluate the effectiveness of such bilingual information.

References

K. W. Church and P. Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29.

I. Dagan, A. Itai, and U. Schwall. 1991. Two languages are more informative than one. In *Proc. of the 29th Annual Meeting of ACL*, pages 130–137.

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

Y. Matsumoto, H. Ishimoto, and T. Utsuro. 1993. Structural matching of parallel texts. In *Proc. of the 31st Annual Meeting of ACL*, pages 23 – 30.

NLRI, (National Language Research Institute). 1993. *Word List by Semantic Principles.* Syuei Syuppan. (in Japanese).

P. Resnik. 1992. WordNet and distributional analysis: A class-based approach to lexical discovery. In *Proc. of the AAAI-92 Workshop on Statistically-Based Natural Language Programming Techniques*, pages 48–56.

S. R. Roget. 1911. *Roget's Thesaurus.* Crowell Co.

M. Shimizu and S. Narita, editors. 1979. *Japanese-English Dictionary.* Kodansha Gakujutsu Bunko.

T. Utsuro, Y. Matsumoto, and M. Nagao. 1993. Verbal case frame acquisition from bilingual corpora. In *Proc. of the 13th IJCAI*, pages 1150–1156.

T. Utsuro. 1995. Class-based sense classification of verbal polysemy in case frame acquisition from parallel corpora. In *Proc. of the 3rd Natural Language Processing Pacific Rim Symposium*, pages 671–677.