Classifying Japanese Polysemous Verbs based on Fuzzy C-means Clustering

Yoshimi Suzuki
Interdisciplinary Graduate School of Medicine and Engineering
University of Yamanashi, Japan
ysuzuki@yamanashi.ac.jp

Fumiyo Fukumoto
Interdisciplinary Graduate School of Medicine and Engineering
University of Yamanashi, Japan
fukumoto@yamanashi.ac.jp

Abstract

This paper presents a method for classifying Japanese polysemous verbs using an algorithm to identify overlapping nodes with more than one cluster. The algorithm is a graph-based unsupervised clustering algorithm, which combines a generalized modularity function, spectral mapping, and fuzzy clustering technique. The modularity function for measuring cluster structure is calculated based on the frequency distributions over verb frames with selectional preferences. Evaluations are made on two sets of verbs including polysemies.

1 Introduction

There has been quite a lot of research concerned with automatic clustering of semantically similar words or automatic retrieval of collocations among them from corpora. Most of this work is based on similarity measures derived from the distribution of words in corpora. However, the facts that a single word does have more than one sense and that the distribution of a word in a corpus is a mixture of usages of different senses of the same word often hamper such attempts. In general, restriction of the subject domain makes the problem of polysemy less problematic. However, even in texts from a restricted domain such as *economics* or *sports*, one encounters quite a large number of polysemous words. Therefore, semantic classification of polysemies has been an interest since the earliest days when a number of large scale corpora have become available.

In this paper, we focus on Japanese polysemous verbs, and present a method for polysemous verb classification. We used a graph-based unsupervised clustering algorithm (Zhang, 2007). The algorithm combines the idea of modularity function Q, spectral relaxation and fuzzy c-means clustering method to identify overlapping nodes with more than one cluster. The modularity function measures the quality of a cluster structure. Spectral mapping performs a dimensionality reduction which makes it possible to cluster in the very high dimensional spaces. The fuzzy c-means allows for the detection of nodes with more than one cluster. We applied the algorithm to cluster polysemous verbs. The modularity function for measuring the quality of a cluster structure is calculated based on the frequency distributions over verb frames with selectional preferences. We collected semantic classes from IPAL Japanese dictionary (IPAL, 1987), and used them as a gold standard data. IPAL lists about 900 Japanese basic verbs, and categorizes each verb into multiple senses. Moreover, the categorization is based on verbal syntax with respect to the choice of its arguments. Therefore, if the clustering algorithm induces a polysemous verb classification on the basis of verbal syntax, then the resulting classification should agree the IPAL classes. We used a large Japanese newspaper corpus and EDR (Electronic Dictionary Research) dictionary (EDR, 1986) to obtain verbs and their subcategorization frames with selectional preferences [1]. The results obtained using two data sets were better than the baseline, EM algorithm.

The rest of the paper is organized as follows. The next section presents related work. After describing Japanese verb with selectional preferences, we present a distributional similarity in Section 4, and a graph-based unsupervised clustering algorithm in Section 5. Results using two data sets are reported in Section 6. We give our conclusion in Section 7.

[1] We did not use IPAL, but instead EDR sense dictionary. Because IPAL did not have senses for the case filler which were used to create selectional preferences.

Proceedings of the 2009 Workshop on Graph-based Methods for Natural Language Processing, ACL-IJCNLP 2009, pages 32–40,
Suntec, Singapore, 7 August 2009. ©2009 ACL and AFNLP

2 Related Work

Graph-based algorithms have been widely used to classify semantically similar words (Jannink, 1999; Galley, 2003; Widdows, 2002; Muller, 2006). Sinha and Mihalcea proposed a graph-based algorithm for unsupervised word sense disambiguation which combines several semantic similarity measures including Resnik's metric (Resnik, 1995), and algorithms for graph centrality (Sinha, 2007). They reported that the results using the SENSEVAL-2 and SENSEVAL-3 English all-words data sets lead to relative error rate reductions of 5 - 8% as compared to the previsous work (Mihalcea, 2005). More recently, Matsuo *et al*. (2006) presented a method of word clustering based on Web counts using a search engine. They applied $Newman$ clustering (Newman, 2004) for identifying word clusters. They reported that the results obtained by the algorithm were better than those obtained by average-link agglomerative clustering using 90 Japanese noun words. However, their method relied on hard-clustering models, and thus have largely ignored the issue of polysemy that word belongs to more than one cluster.

In contrast to hard-clustering algorithms, soft clustering allows that words to belong to more than one cluster. Much of the previous work on word classification with soft clustering is based on the EM algorithm (Pereira, 1993). Torisawa *et al.*, (2002) presented a method to detect associative relationships between verb phrases. They used the EM algorithm to calculate the likelihood of co-occurrences, and reported that the EM is effective to produce associative relationships with a certain accuracy. More recent work in this direction is that of Schulte *et al.*, (2008). They proposed a method for semantic verb classification based on verb frames with selectional preferences. They combined the EM training with the MDL principle. The MDL principle is used to induce WordNet-based selectional preferences for arguments within subcategorization frames. The results showed the effectiveness of the method. Our work is similar to their method in the use of verb frames with selectional preferences. Korhonen *et al*. (2003) used verb–frame pairs to cluster verbs into Levin-style semantic classes (Korhonen, 2003). They used the Information Bottleneck, and classified 110 test verbs into Levin-style classes. They had a focus on the interpretation of verbal polysemy as represented by the soft clusters: they interpreted polysemy as multiple-hard assignments.

In the context of Japanese taxonomy of verbs and their classes, Utsuro *et al*. (1995) proposed a class-based method for sense classification of verbal polysemy in case frame acquisition from parallel corpora (Utsuro, 1995). A measure of bilingual class/class association is introduced and used for discovering sense clusters in the sense distribution of English predicates and Japanese case element nouns. They used the test data consisting of 10 English and Japanese verbs taken from Roget's Thesaurus and BGH (Bunrui Goi Hyo) (BGH, 1989). They reported 92.8% of the discovered clusters were correct. Tokunaga *et al*. (1997) presented a method for extending an existing thesaurus by classifying new words in terms of that thesaurus. New words are classified on the basis of relative probabilities of a word belonging to a given word class, with the probabilities calculated using noun-verb co-occurrence pairs. Experiments using the Japanese BGH thesaurus showed that new words can be classified correctly with a maximum accuracy of more than 80%, while they did not report in detail whether the clusters captured polysemies.

3 Selectional Preferences

A major approach on word clustering task is to use distribution of a word in a corpus, *i.e.*, words are classified into classes based on their distributional similarity. Similarity measures based on distributional hypothesis compare a pair of weighted feature vectors that characterize two words (Hindle, 1990; Lin, 1998; Dagan, 1999).

Like previous work on verb classification, we used subcategorization frame distributions with selectional preferences to calculate similarity between verbs (Schulte, 2008). We used the EDR dictionary of selectional preferences consisting of 5,269 basic Japanese verbs and the EDR concept dictionary (EDR, 1986). For selectional preferences, the dictionary has each concept of a verb, the group of possible co-occurrence surface-level case particles, the types of concept relation label that correspond to the surface-level case as well as the range of possible concepts that may fill the deep-level case. Figure 1 illustrates an example of a verb "*taberu* (eat)".

In Figure 1, "Sentence pattern" refers to the co-occurrence pattern between a verb and a noun

[Sentence pattern]	<word1> *ga*	<word2> *wo*	*taberu* (eat)
[Sense relation]	agent	object	
[Case particle]	*ga* (nominative)	*wo* (accusative)	
[Sense identifier]	30f6b0 (human);30f6bf (animal)	30f6bf(animal);30f6ca(plants); 30f6e5(parts of plants); 3f9639(food and drink); 3f963a(feed)	

Figure 1: An example of a verb "*taberu* (eat)"

with a case marker. "Sense relation" expresses the deep-level case, while "Case particle" shows the surface-level case. "Sense identifier" refers to the range of possible concepts for the case filler. The subcategorization frame pattern of a sentence (1), for example consists of two arguments with selectional preferences and is given below:

(1) *Nana_ga apple_wo taberu.*
'Nana eats an apple.'

taberu	30f6b0_*ga*	3f9639_*wo*
eat	human_nom	entity_acc

In the above frame pattern, x of the argument "x_y" refers to sense identifier and y denotes case particle.

4 Distributional Similarity

Various similarity measures have been proposed and used for NLP tasks (Korhonen, 2002). In this paper, we concentrate on three distance-based, and entropy-based similarity measures. In the following formulae, x and y refer to the verb vectors, their subscripts to the verb subcategorization frame values.

1. **The Cosine measure (Cos):** The cosine measures the similarity of the two vectors x and y by calculating the cosine of the angle between vectors, where each dimension of the vector corresponds to each frame with selectional preferences patterns of verbs and each value of the dimension is the frequency of each pattern.

2. **The Cosine measure based on probability of relative frequencies (rfCos):** The differences between the cosine and the value based on relative frequencies of verb frames with selectional preferences are the values of each dimension, *i.e.*, the former are frequencies of each pattern and the latter are the fraction of the total number of verb frame patterns belonging to the verb.

3. L_1 **Norm (L_1):** The L_1 Norm is a member of a family of measures known as the Minkowski Distance, for measuring the distance between two points in space. The L_1 distance between two verbs can be written as:

$$L_1(x,y) \quad = \quad \sum_{i=1}^{n} \mid x_i - y_i \mid .$$

4. **Kullback-Leibler (*KL*):** Kullback-Leibler is a measure from information theory that determines the inefficiency of assuming a model probability distribution given the true distribution.

$$KL(x,y) \quad = \quad \sum_{i=1}^{n} P(x_i) * \log \frac{P(x_i)}{P(y_i)}.$$

where $P(x_i) = \frac{x_i}{|x|}$. *KL* is not defined in case $y_i = 0$. So, the probability distributions must be smoothed (Korhonen, 2002). We used two smoothing methods, *i.e.*, Add-one smoothing and Witten and Bell smoothing (Witten, 1991).[2] Moreover, two variants of *KL*, α-skew divergence and the Jensen-Shannon, were used to perform smoothing.

5. **α-skew divergence (α div.):** The α-skew divergence measure is a variant of *KL*, and is defined as:

$$\alpha div(x,y) \quad = \quad KL(y, \alpha \cdot x + (1-\alpha) \cdot y).$$

Lee (1999) reported the best results with α = 0.9. We used the same value.

6. **The Jensen-Shannon (*JS*):** The Jensen-Shannon is a measure that relies on the assumption that if x and y are similar, they are close to their average. It is defined as:

[2]We report Add-one smoothing results in the evaluation, as it was better than Witten and Bell smoothing.

$$JS(x,y) = \frac{1}{2}[KL(x, \frac{x+y}{2}) + KL(y, \frac{x+y}{2})].$$

All measures except Cos and rfCos showed that smaller values indicate a closer relation between two verbs. Thus, we used inverse of each value.

5 Clustering Method

The clustering algorithm used in this study was a graph-based unsupervised clustering reported by (Zhang, 2007). This algorithm detects overlapping nodes by the combination of a modularity function based on Newman Girvan's Q function (Newman, 2004), spectral mapping that maps input nodes into Euclidean space, and fuzzy c-means clustering which allows node to belong to more than one cluster. They evaluated their method by applying several data including the American college football team network, and found that the algorithm successfully detected overlapping nodes. We thus used the algorithm to cluster verbs.

Here are the key steps of the algorithm: Given a set of input verbs $V = \{v_1, v_2, \cdots v_n\}$, an upper bound K of the number of clusters, the adjacent matrix $A = (a_{ij})_{n \times n}$ of an input verbs and a threshold λ that can convert a soft assignment into final clustering, *i.e.*, the value of λ decreases, each verb is distributed into larger number of clusters. We calculated the adjacent matrix A by using one of the similarity measures mentioned in Section 4, *i.e.*, the value of the edge between v_i and v_j. a_{ij} refers to the similarity value between them.

1. Form a diagonal matrix $D = (d_{ii})$, where $d_{ii} = \sum_k a_{ik}$.

2. Form the eigenvector matrix $E_K = [e_1, e_2, \cdots, e_K]$ by calculating the top K eigenvectors of the generalized eigensystem $Ax = tDx$.

3. For each value of k, $2 \leq k \leq K$:

   (a) Form the matrix $E_k = [e_2, \cdots, e_k]$ where e_k refers to the top k-th eigenvector.

   (b) Normalize the rows of E_k to unit length using Euclidean distance norm.

   (c) Cluster the row vectors of E_k using fuzzy c-means to obtain a soft assignment matrix U_k. Fuzzy c-means is carried out through an iterative optimization (minimization) of the objective function J_m with the update of membership degree u_{ij} and the cluster centers c_j. J_m is defined as:

$$J_m = \sum_{i=1}^{n}\sum_{j=1}^{k} u_{ij}^m \mid\mid v_i - c_j \mid\mid^2,$$

   where u_{ij} is the membership degree of v_i in the cluster j, and $\sum_j u_{ij} = 1$. $m \in [1, \infty]$ is a weight exponent controlling the degree of fuzzification. c_j is the d-dimensional center of the cluster j.
   $\mid\mid v_i - c_j \mid\mid$ is defined as:

$$\mid\mid v_i - c_j \mid\mid^2 = (v_i - c_j)E(v_i - c_j)^T.$$

   where E denotes an unit matrix. The procedure converges to a saddle point of J_m.

4. Pick the k and the corresponding $n \times k$ soft assignment matrix U_k that maximizes the modularity function $\tilde{Q}(U_k)$. Here $U_k = [u_1, \cdots u_k]$ with $0 \leq u_{ic} \leq 1$ for each $c = 1, \cdots, k$, and $\sum_1^k u_{ic} = 1$ for each $i = 1, \cdots, n$. A modularity function of a soft assignment matrix is defined as:

$$\tilde{Q}(U_k) = \sum_{c=1}^{k}[\frac{A(\tilde{V}_c, \tilde{V}_c)}{A(V,V)} - (\frac{A(\tilde{V}_c, V)}{A(V,V)})^2],$$

   where

$$\begin{aligned} A(\tilde{V}_c, \tilde{V}_c) &= \sum_{i \in \tilde{V}_c, j \in \tilde{V}_c} \{\frac{(u_{ic} + u_{jc})}{2}\} a_{ij}, \\ A(\tilde{V}_c, V) &= A(\tilde{V}_c, \tilde{V}_c) + \\ &\quad \sum_{i \in \tilde{V}_c, j \in V \setminus \tilde{V}_c} \{\frac{(u_{ic} + (1 - u_{jc}))}{2}\} a_{ij}, \\ A(V,V) &= \sum_{i \in V, j \in V} a_{ij}. \end{aligned}$$

   $\tilde{Q}(U_k)$ shows comparison of the actual values of internal or external edges with its respective expectation value under the assumption of equally probable links and given data sizes.

6 Experiments

6.1 Experimental setup

We created test verbs using two sets of Japanese Mainichi newspaper corpus. One is a set consisting one year (2007) newspapers (We call it a set from 2007), and another is a set of 17 years (from 1991 to 2007) Japanese Mainichi newspapers (We call it a set from 1991_2007). For each set, all Japanese documents were parsed using the syntactic analyzer Cabocha (Kudo, 2003). We selected verbs, each frequency $f(v)$ is, $500 \leq f(v) \leq 10{,}000$. As a result, we obtained 279 verbs for a set from 2007 and 1,692 verbs for a set from 1991_2007. From these verbs, we chose verbs which appeared in the machine readable dictionary, IPAL. This selection resulted in a total of 81 verbs for a set from 2007, and 170 verbs, for a set from 1991_2007. We obtained Japanese verb frames with selectional preferences using these two sets. We extracted sentence patterns with their frequencies. Noun words within each sentence were tagged sense identifier by using the EDR Japanese sense dictionary. As a result, we obtained 56,400 verb frame patterns for a set from 2007, and 300,993 patterns for a set from 1991_2007.

We created the gold standard data, verb classes, using IPAL. IPAL lists about 900 Japanese verbs and categorizes each verb into multiple senses, based on verbal syntax and semantics. It also listed *synonym* verbs. Table 1 shows a fragment of the entry associated with the Japanese verb *taberu*. The verb "*taberu*" has two senses, "eat" and "live". "pattern" refers to the case frame(s) associated with each verb sense. According to the IPAL, we obtained verb classes, each class corresponds to a sense of each verb. There are 87 classes for a set from 2007, and 152 classes for a set from 1991_2007. The examples of the test verbs and their senses are shown in Table 2.

For evaluation of verb classification, we used the precision, recall, and F-score, which were defined by (Schulte, 2000), especially to capture how many verbs does the algorithm actually detect more than just the predominant sense.

For comparison against polysemies, we utilized the EM algorithm which is widely used as a soft clustering technique (Schulte, 2008). We followed the method presented in (Rooth, 1999). We used a probability distribution over verb frames with selectional preferences. The initial probabilities were often determined randomly. We set the initial probabilities by using the result of the standard k-means. For k-means, we used 50 random replications of the initialization, each time initializing the cluster center with k randomly chosen. We used up to 20 iterations to learn the model probabilities.

Table 3: Results for a set from 2007

Method	m	λ	C	Prec	Rec	F
FCM	2.0	0.09	74	.815	.483	**.606**
FCM(none)	1.5	0.07	74	.700	.477	.567
EM	–	–	87	.308	.903	.463

Table 4: Results against each measure

Measure	m	λ	C	Prec	Rec	F
cos	3.0	0.02	74	.660	.517	.580
rfcos	2.0	0.04	74	.701	.488	.576
L_1	2.0	0.04	74	.680	.500	.576
KL	2.0	0.09	74	.815	.483	.606
α div.	2.0	0.04	74	.841	.471	.604
JS	1.5	0.03	74	.804	.483	.603
EM	–	–	87	.308	.903	.463

6.2 Basic results

The results using a set from 2007 are shown in Table 3. We used *KL* as a similarity measure in FCM. "FCM(none)" shows the result not applying a spectral mapping, *i.e.*, we applied fuzzy c-means to each vector of verb, where each dimension of the vector corresponds to each frame with selectional preferences. "m" and "λ" refer to the parameters used by Fuzzy C-means. "C" refers to the number of clusters obtained by each method. "m", "λ" and "C" in Table 3 denote the value that maximized the F-score. "C" in the EM is fixed in advance. The result of EM shows the best result among 20 iterations. As can be seen clearly from Table 3, the result obtained by fuzzy c-means was better to the result by EM algorithm. Table 3 also shows that a dimensionality reduction, *i.e.*, spectral mapping improved overall performance, especially we have obtained better precision. The result suggests that a dimensionality reduction is effective for clustering. Table 4 shows the results obtained by using each similarity measure. As we can see from Table 4, the overall results obtained by information theory based measures, *KL*, α div., and *JS* were slightly better to the results obtained by other distance based measures.

We note that the fuzzy c-means has two parameters λ and m, where λ is a threshold of the as-

Table 1: A fragment of the entry associated with the Japanese verb "*taberu*"

Sense_id	Pattern		Synonyms
1	*kare*(he)_*ga*(nominative)	*soba*(noodles)_*wo*(accusative)	*kuu* (eat)
2	*kare* (he)_*ga*(nominative)	*fukugyo*(a part-time job)_*de*(accusative)	*kurasu* (live)

Table 2: Examples of test verbs and their polysemic gold standard senses

Id	Sense	Verb Classes	Id	Sense	Verb Classes
1	treat	$\{ashirau, atsukau\}$	11	tell	$\{oshieru, shimesu, shiraseru\}$
2	prey	$\{negau, inoru\}$	12	persuade	$\{oshieru, satosu\}$
3	wish	$\{negau, nozomu\}$	13	congratulate	$\{iwau, syukufukusuru\}$
4	ask	$\{negau, tanomu\}$	14	accept	$\{uketoru, ukeru, morau, osameru\}$
5	leave	$\{saru, hanareru\}$	15	take	$\{uketoru, toru, kaisyakusuru, miru\}$
6	move	$\{saru, utsuru\}$	16	lose	$\{ushinau, nakusu\}$
7	pass	$\{saru, kieru, sugiru\}$	17	miss	$\{ushinau, torinogasu, itusuru\}$
8	go	$\{saru, sugiru, iku\}$	18	survive, lose	$\{ushinau, nakusu, shinareru\}$
9	remove	$\{saru, hanareru, toozakeru$ $torinozoku\}$	19	give	$\{kubaru, watasu, wakeru\}$
10	lead	$\{oshieru, michibiku, tugeru\}$	20	arrange	$\{kubaru, haichisuru\}$

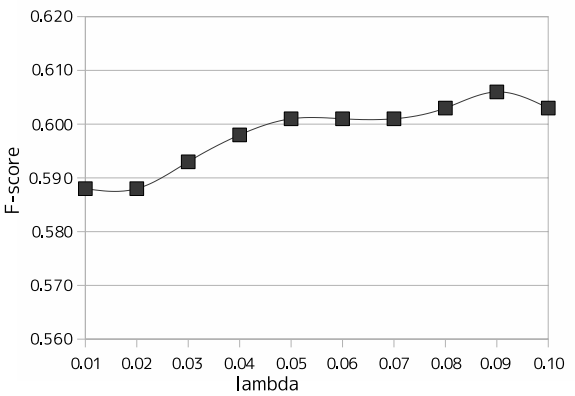

Figure 2: F-score against λ

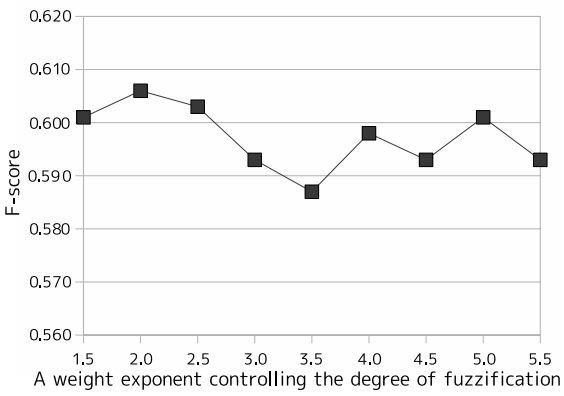

Figure 3: F-score against m

signment in the fuzzy c-means, and m is a weight controlling the degree of fuzzification. To examine how these parameters affect the overall performance of the algorithm, we performed experiments by varying these parameters. Figure 2 illustrates F-score of polysemies against the value of λ. We used *KL* as a similarity measure, m = 2, and C = 74.

As shown in Figure 2, the best result was obtained when the value of λ was 0.09. When λ value was larger than 0.09, the overall performance decreased, and when it exceeded 1.2, no verbs were assigned to multiple sense. Figure 3 illustrates F-score against the value of m. As illustrated in Figure 3, we could not find effects on accuracy against the value of m. It is necessary to investigate on the influence of the parameter m by performing further quantitative evaluation.

6.3 Error analysis against polysemy

We examined whether 46 polysemous verbs in a set from 2007 were correctly classified into classes. We manually analyzed clustering results obtained by running fuzzy c-means with *KL* as a similarity measure. They were classified into three types of error.

1. **Partially correct:** Some senses of a polysemous verb were correctly identified, but others were not. The first example of this pattern is that "*nigiru*" has at least two senses, "*motsu* (have)" and "*musubu* (double)". However, only one sense was identified correctly. The second example is that one of the senses of the verb "*watasu*" was classified correctly into the class "*ataeru* (give)", while it was classified incorrectly into the class "*uru* (sell)". This was the most frequent error type.

{*nigiru*, *motsu* (have)}
ϕ

{*watasu*, *ataeru* (give)}
{*watasu*, *uru* (sell)}

2. **Polysemous verbs classified into only one cluster:** "*hakobu*" has two senses "carry", and "progress". However, it was classified into one cluster including verbs "*motuteiku* (carry)", and "*susumu* (progress)". Because it often takes the same nominative subjects such as "human" and accusative object such as "abstract".

{*hakobu* (carry, progress),
motuteiku (carry), *susumu* (progress)}

3. **Polysemous verb incorrectly classified into clusters:** The polysemous verb "*hataraku*" has two senses, "work", and "operate". However, it was classified incorrectly into "*ochiru* (fall)" and "*tsukuru* (make)".

{*hataraku* (work, operate), *ochiru* (fall),
tsukuru (make)}

Apart from the above error analysis, we found that we should improve the definition and demarcation of semantic classes by using other existing thesaurus, *e.g.*, EDR or BGH (Bunrui Goi Hyo) (BGH, 1989). We recall that we created the gold standard data by using synonymous information. However, the algorithm classified some antonymous words such as "*uketoru*" (receive) and "*watasu*" (give) into one cluster. Similarly, transitive and intransitive verbs are classified into the same cluster. For example, intransitive verb of the verb "*ochiru*" (drop) is "*otosu*". They were classified into one cluster. It would provide further potential, *i.e.*, not only to improve the accuracy of classification, but also to reveal the relationship between semantic verb classes and their syntactic behaviors.

An investigation of the resulting clusters revealed another interesting direction of the method. We found that some senses of a polysemous verb which is not listed in the IPAL are correctly identified by the algorithm. For example, "*ukeireru*" and "*yurusu*" (forgive) were correctly classified into one cluster. Figure 4 illustrates a sample of verb frames with selectional preferences extracted by our method.

"*ukeireru*" and "*yurusu*" in Table 4 have the same frame pattern, and the sense identifiers of the case filler "*wo*", for example, are "a human being" (0f0157) and "human" (30f6b0). However, these verbs are not classified into one class in the IPAL: "*ukeireru*" is not listed in the IPAL as a synonym verb of "*yurusu*". The example illustrates that these verbs within a cluster are semantically related, and that they share obvious verb frames with intuitively plausible selectional preferences. This indicates that we can extend the algorithm to solve this resource scarcity problem: semantic classification of words which do not appear in the resource, but appear in corpora.

Table 5: Results for a set from 1991_2007

Method	m	λ	C	Prec	Rec	F
FCM	2.0	0.24	152	.792	.477	**.595**
FCM(none)	2.0	0.07	147	.687	.459	.550
EM	–	–	152	.284	.722	.408

6.4 Results for a set of verbs from 1991_2007 corpus

One goal of this work was to develop a clustering methodology with respect to the automatic recognition of Japanese verbal polysemies covering large-scale corpora. For this task, we tested a set of 170 verbs including 82 polysemies. The results are shown in Table 5. We used *KL* as a similarity measure in FCM. Each value of the parameter shows the value that maximized the F-score.

As shown in Table 5, the result obtained by fuzzy c-means was as good as for the smaller set, a set of 78 verbs. Moreover, we can see that the fuzzy c-means is better than the EM algorithm and the method not applying a spectral mapping, as an increase in the F-score of 18.7% compared with the EM, and 4.5% compared with a method without spectral mapping. This shows that our method is effective for a size of the input test data consisting 178 verbs.

One thing should be noted is that when the algorithm is applied to large data, it is computationally expensive. There are at least two ways to address the problem. One is to use several methods

[Sentence pattern]	<word1> ga	<word2> wo	$ukeireru$ / $yurusu$ (forgive)
[Concept relation]	agent	object	
[Case particle]	ga (nominative)	wo (accusative)	
[Sense identifier]	0ee0de; 0f58b4; 0f98ee	0f0157; 30f6b0	

0ee0de: the part of a something written that makes reference to a particular matter
0f58b4: a generally-held opinion
0f98ee: the people who citizens of a nation
0f0157: a human being
30f6b0: human

Figure 4: Extracted Verb frames of "$ukeireru$" and "$yurusu$" (forgive)

of fuzzy c-means acceleration. Kelen *et al*. (2002) presented an efficient implementation of the fuzzy c-means algorithm, and showed that the algorithm had the worse-case complexity of $O(nK^2)$, where n is the number of nodes, and K is the number of eigenvectors. Another approach is to parallelize the algorithm by using the Message Passing Interface (MPI) to estimate the optimal number of k ($2 \leq k \leq K$). This is definitely worth trying with our method.

7 Conclusion

We have developed an approach for classifying Japanese polysemous verbs using fuzzy c-means clustering. The results were comparable to other unsupervised techniques. Future work will assess by a comparison against other existing soft clustering algorithms such as the Clique Percolation method (Palla, 2005). Moreover, it is necessary to apply the method to other verbs for quantitative evaluation. New words including polysemies are generated daily. We believe that classifying these words into semantic classes potentially enhances many semantic-oriented NLP applications. It is necessary to apply the method to other verbs, especially low frequency of verbs to verify that claim.

Acknowledgments

This work was supported by the Grant-in-aid for the Japan Society for the Promotion of Science (JSPS).

References

E. Iwabuchi. 1989. Word List by Semantic Principles, *National Language Research Institute Publications*, Shuei Shuppan.

I. Dagan and L. Lee and F. C. N. Pereira. 1999. Similarity-based Models of Word Cooccurrence Probabilities. *Machine Learning*, 34(1-3), pages 43–69.

Japan Electronic Dictionary Research Institute, Ltd. http://www2.nict.go.jp/r/r312/EDR/index.html

M. Galley and K. McKeown. 2003. Improving Word Sense Disambiguation in Lexical Chaining, In *Proc. of 19th International Joint Conference on Artificial Intelligence*, pages 1486–1488.

D. Hindle. 1990. Noun Classification from Predicate-Argument Structures, In *Proc. of 28th Annual Meeting of the Association for Computational Linguistics*, pages 268–275.

GSK2007-D. http://www.gsk.or.jp/catalog/GSK2007-D/catalog.html

J. Jannink and G. Wiederhold. 1999. Thesaurus Entry Extraction from an On-line Dictionary, In *Proc. of Fusion'99*.

J. F. Kelen and T. Hutcheson. 2002. Reducing the Time Complexity of the Fuzzy C-means Algorithm, In *Trans. of IEEE Fuzzy Systems*, 10(2), pages 263–267.

A. Korhonen and Y. Krymolowski. 2002. On the Robustness of Entropy-based Similarity Measures in Evaluation of Subcategorization Acquisiton Systems. In *Proc. of the 6th Conference on Natural Language Learning*, pages 91–97.

A. Korhonen and Y. Krymolowski and Z. Marx. 2003. Clustering Polysemic Subcategorization Frame Distributions Semantically. In *Proc. of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 64–71.

T.Kudo and Y.Matsumoto. 2003. Fast Methods for Kernel-based Text Analysis. In *Proc. of 41th ACL*, pages 24–31.

L. Lee. 1999. Measures of Distributional Similarity. In *Proc. of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 25–32.

D. Lin. 1998. Automatic Retrieval and Clustering of Similar Words, In *Proc. of 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*, pages 768–773.

Y. Matsuo and T. Sakaki and K. Uchiyama and M. Ishizuka. 2006. Graph-based Word Clustering using a Web Search Engine, In *Proc. of 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP2006)*, pages 542–550.

R. Mihalcea. 2005. Unsupervised Large Vocabulary Word Sense Disambiguation with Graph-based Algorithms for Sequence Data Labeling, In *Proc. of the Human Language Technology / Empirical Methods in Natural Language Processing Conference*, pages 411–418.

P. Muller and N. Hathout and B. Gaume. 2006. Synonym Extraction Using a Semantic Distance on a Dictionary, In *Proc. of the Workshop on TextGraphs*, pages 65–72.

M.E.J.Newman. 2004. Fast Algorithm for Detecting Community Structure in Networks, *Physical Review*, E 2004, 69, 066133.

G. Palla and I. Derényi and I. Farkas and T. Vicsek. 2005. Uncovering the Overlapping Community Structure of Complex Networks in Nature and Society, *Nature*. 435(7043), 814–8.

F. Pereira and N. Tishby and L. Lee. 1993. Distributional Clustering of English Words. In *Proc. of the 31st Annual Meeting of the Association for Computational Linguistics*, pages 183–190.

P. Resnik. 1995. Using Information Content to Evaluate Semantic Similarity in a Taxonomy. In *Proc. of 14th International Joint Conference on Artificial Intelligence*, pages 448–453.

M. Rooth et al. 1999. Inducing a Semantically Annotated Lexicon via EM-Based Clustering, In *Proc. of 37th ACL*, pages 104–111.

R. Sinha and R. Mihalcea. 2007. Unsupervised Graph-based Word Sense Disambiguation Using Measures of Word Semantic Similarity. In *Proc. of the IEEE International Conference on Semantic Computing*, pages 46–54.

S. Schulte im Walde. 2000. Clustering Verbs Semantically according to their Alternation Behaviour. In *Proc. of the 18th COLING*, pages 747–753.

S. Schulte im Walde et al. 2008. Combining EM Training and the MDL Principle for an Automatic Verb Classification Incorporating Selectional Preferences. In *Proc. of the 46th ACL*, pages 496–504.

T. Tokunaga and A. Fujii and M. Iwayama and N. Sakurai and H. Tanaka. 1997. Extending a thesaurus by classifying words. In *Proc. of the ACL-EACL Workshop on Automatic Information Extraction and Building of Lexical Semantic Resources*, pages 16–21.

K. Torisawa. 2002. An Unsupervised Learning Method for Associative Relationships between Verb Phrases, In *Proc. of 19th International Conference on Computational Linguistics (COLING2002)*, pages 1009–1015.

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.

D. Widdows and B. Dorow. 2002. A Graph Model for Unsupervised Lexical Acquisition. In *Proc. of 19th International conference on Computational Linguistics (COLING2002)*, pages 1093–1099.

I. H. Witten and T. C. Bell. 1991. The Zero-Frequency Problem: Estimating the Probabilities of Novel Events in Adaptive Text Compression. *IEEE Transactions on Information Theory*, 37(4), pages 1085–1094.

S. Zhang et al. 2007. Identification of Overlapping Community Structure in Complex Networks using Fuzzy C-means Clustering. *PHYSICA* A, 374, pages 483–490.