IEICE TRANS. INF. & SYST., VOL.E92–D, NO.12 DECEMBER 2009

PAPER Special Section on Natural Language Processing and its Applications

Co-clustering with Recursive Elimination for Verb Synonym Extraction from Large Text Corpus

Koichi TAKEUCHI[†a)], *Member and* **Hideyuki TAKAHASHI**[†], *Nonmember*

SUMMARY The extraction of verb synonyms is a key technology to build a verb dictionary as a language resource. This paper presents a co-clustering-based verb synonym extraction approach that increases the number of extracted meanings of polysemous verbs from a large text corpus. For verb synonym extraction with a clustering approach dealing with polysemous verbs can be one problem issue because each polysemous verb should be categorized into different clusters depending on each meaning; thus there is a high possibility of failing to extract some of the meanings of polysemous verbs. Our proposed approach can extract the different meanings of polysemous verbs by recursively eliminating the extracted clusters from the initial data set. The experimental results of verb synonym extraction show that the proposed approach increases the correct verb clusters by about 50% with a 0.9% increase in precision and a 1.5% increase in recall over the previous approach.

key words: *verb synonyms, co-clustering, polysemy, recursive elimination*

1. Introduction

Since a verb's meaning is highly dependent on the nouns that the verb takes as arguments verb synonyms with a shared meaning can be extracted by considering the dependency relations between verbs and nouns. For example, the meaning of *employ* expressed in "employ a person" is the same of "hire" while the other meaning of *employ* in "employ an approach" equals "use." Since each meaning of a verb takes a different set of argument nouns: e.g., "employ/hire" takes "person/him" as arguments while "employ/use" takes different arguments; the "method/approach" a co-clustering approach that clusters both verbs and nouns is appropriate for verb-clustering tasks.

Previous work shows the possibility of applying Aizawa's co-clustering approach into verb synonym extraction for Japanese from large text corpus. Aizawa's co-clustering is a soft clustering approach that can categorize one word into several clusters; however, some meanings of verbs are submerged in the other clusters because of difficulty of maximizing the extracted clusters.

In this paper we therefore propose a revised co-clustering approach that can actively extract different meanings of polysemous verbs by recursively eliminating the extracted meaning. Experimental results show that the proposed approach outperforms the previous approach as well as a single-clustering approach based on PLSI. We also confirm the characteristics of the proposed approach that depends on the quality or the scale of the input text corpus. From the results we will show that the proposed approach is highly promising for verb synonym extraction.

Manuscript received March 20, 2009.
Manuscript revised July 17, 2009.

[†]The authors are with the Graduate School of Natural Science and Technology, Okayama University, Okayama-shi, 700–8530 Japan.
a) E-mail: koichi@cl.cs.okayama-u.ac.jp

DOI: 10.1587/transinf.E92.D.2334

2. Background Issues

The major word clustering schemes are word distribution-based methods [1], [2] and decomposition-based methods [3]–[5]. The decompositional methods cluster words on the basis of the orthogonal vectors of the assumed latent semantic clusters between words. Their aim is to make compact orthogonal vectors based on global optimization such as the EM algorithm; however, they do not seem suitable for the extraction of verb synonyms because both verbs and their arguments are polysemous and thus simultaneous clustering such as the co-clustering is more feasible for the extraction of verb synonyms.

Recently several co-clustering approaches have been proposed, such as the Dirichlet process-based approaches [6], [7], the graph partition-based approaches [8], and the mutual information-based approaches [9]. These approaches categorize all of the input data into some clusters; however, since a large amount of input data usually contains noise that is inappropriate for clustering, and thus these approaches do not seem appropriate for application to large-scale data. We applied Dhillon's co-clustering tools to our co-occurrence data based on verb and noun pairs with a case extracted from one year of Japanese newspaper articles, but we found that the co-clustering tools did not work for only a one-year corpus. Additionally, we found that Kurihara's co-clustering approach was not available for this verb-clustering task because of the limitation of output cluster numbers that it can deal with[*]. In contrast to these approaches, Aizawa's co-clustering approach can be considered a feasible approach for the verb-clustering tasks because of the following characteristics: (1) Extraction of only tightly associated substructures from the input data instead of categorizing all input data into some clusters indicating that this approach will be robust for noisy input data. (2) Use of entropy-based criteria for cluster evaluation so that the generated clusters can be determined independently of other clusters to avoid an explosive amount of calculation. Even though these two characteristics are effective for large-scale data they cause a local minimum problem: some meanings of

[*]The limitation is due to their implementation.

polysemous verbs are submerged in other clusters due to the hill-climbing algorithm.

In this paper we thus propose a simple but an effective approach that actively extracts other meanings from the extracted meaning of the polysemous verb by recursively eliminating the extracted meaning from the initial data. The details are described in Sect. 3.2.

3. Recursive Elimination Approach

The problem of insufficient extraction of verb meanings from the initial data is highly dependent on the detailed steps of Aizawa's approach. Thus first we explain the related part of the detailed approach of verb synonym extraction based on Aizawa's co-clustering approach and clarify why the synonyms remain. Second we explain how to extract the remaining possible synonyms from the clusters utilizing the characteristics of Aizawa's co-clustering approach.

3.1 Aizawa's Co-clustering

The following are the two main steps of Japanese verb synonym extraction based on Aizawa's approach[†]:

(1) Select a starting verb and make an initial bipartite graph that consists of verb and noun pairs with case markers.
(2) If the entropy-based criteria of a subgraph in the initial bipartite graph is positive, then the subgraph is considered as a correct candidate.

Verb synonym clusters are extracted by repeating the above two steps and varying the initial bipartite graph.

As for (1), each link in a bipartite graph indicates a pair between a verb and a noun with a case maker (hereafter noun-case) that appears as a direct dependent relation of the verb in a sentence of the text corpus. Such dependency relations are automatically annotated from the text corpus using a dependency parser. The following is the procedure for making an initial bipartite graph: select a starting verb; add links from it to the noun-cases that appear as dependent relations to the selected verb; add links from each noun-case to the verbs that have direct dependency relations with the noun-case; now an initial bipartite graph has been constructed.

Next for (2), we describe its essential formula. Let t_i and d_j be an i-th verb and a j-th noun-case, respectively. Let S_T and S_D be a candidate subset of verbs and a subset of noun-cases, respectively. A subgraph can then be defined as a combination of S_T and S_D. Since we expect to extract a cluster of verbs and noun-cases that can be considered paraphrasable, we regard a cluster that has dependency links of as many combinations between verbs and noun-cases as an expected cluster. The basic idea of evaluating such a cluster on the basis of an information theoretic view of the retrieval system [11] is that an expected subgraph (S_T, S_D) will bring a gain in the total mutual information described as the following formula[††]:

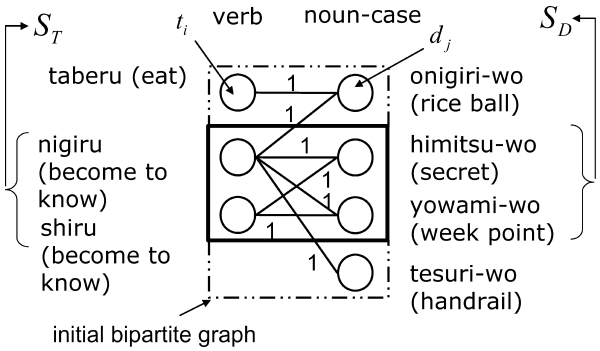

Fig. 1 Example of selecting a correct cluster candidate from an initial bipartite graph. The number at each link denotes the co-occurrence frequency between the verb and the noun-case.

$$\delta I(S_T, S_D) = P(S_T, S_D) \log \frac{P(S_T, S_D)}{P(S_T)P(S_D)} - \sum_{t_i \in S_T} \sum_{d_j \in S_D} P(t_i, d_j) \log \frac{P(t_i, d_j)}{P(t_i)P(d_j)}. \quad (1)$$

If $\delta I(S_T, S_D)$ is positive, this approach will extract the cluster (S_T, S_D) as a correct cluster[†††].

One of the difficulties in extracting clusters by the co-clustering approach is that the above formula does not indicate how to find the best subgraph (S_T, S_D) from the initial bipartite graph[††††], and thus a method is needed that finds promising subgraph candidates. The current method for finding hopeful subgraph candidates is a kind of hill-climbing approach, i.e., making a subgraph candidate by removing element (t_i or d_j) with the least contribution to the initial graph evaluated by Eqs. (2) or (3)[†††††]; and extracting the candidate as a correct subgraph if Eq. (1) is positive for the candidate (Fig. 1).

$$\delta I(t_i, S_D) = \sum_{d_j \in S_D} P(t_i, d_j) \log \frac{P(t_i, d_j)}{P(t_i)P(d_j)}. \quad (2)$$

$$\delta I(S_T, d_j) = \sum_{t_i \in S_T} P(t_i, d_j) \log \frac{P(t_i, d_j)}{P(t_i)P(d_j)}. \quad (3)$$

From the above procedure, the current method will not extract other subgraphs even though other correct subgraphs remain in the initial bipartite graph. Therefore we must develop an extraction procedure without such an explosive amount of calculation.

[†]The details are in Takeuchi [10].
[††]The detailed derivation is described in Aizawa [11].
[†††]Before calculating this formula, we filter out the words with more than 100 links from the bipartite graph because such high occurrence words as "suru" (do) and "aru" (is) can be ignored.
[††††]Checking all combinations of (S_T, S_D) does not seem to be available for large-scale data.
[†††††]Eq. (2) evaluates how verb node t_i contributes to a temporal cluster of (S_T, S_D). If a t_i node contributes to subgraph (S_T, S_D), Eq. (1) will be large, i.e., the second term in the right-hand side of Eq. (1) will be small. Eq. (3)is the same as Eq. (2). The details are described in Aizawa's paper [11].

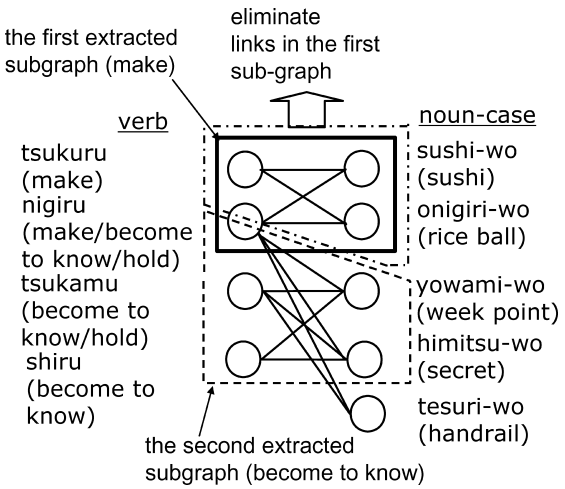

Fig. 2 Possibility of extracting second meanings of polysemous verb "nigiru" (make/become to know/hold).

3.2 Co-clustering with Recursive Elimination

One characteristic of the hill-climbing based approach is that other subgraphs will be extracted if we change the initial state of the bipartite graph. Therefore, another different subgraph will be obtained if we apply the extraction procedure in Sect. 3.1 to the modified bipartite graph from which the extracted clusters were eliminated. This elimination procedure is illustrated in Fig. 2.

In Fig. 2 we assume that verbs "nigiru" (make) and "tsukuru" (make) and noun-cases "sushi-wo" (sushi) and "onigiri-wo" (rice ball) are extracted from the first subgraph of the initial bipartite graph. In this situation, by eliminating the links of the first subgraphs, the second subgraph consists of verbs "nigiru" (become to know), "tsukamu" (become to know), "shiru" (become to know) and noun-cases "yowami-wo" (weakness) and "himitsu-wo" (secret) are extracted. By applying this recursive elimination procedure, the remaining latent subgraphs will be extracted.

4. Experiments of Verb Synonym Extraction

In this section we evaluate the performance of our proposed co-clustering approach with recursive elimination (CCRE) by comparing it with Aizawa's co-clustering approach (i.e., without recursive elimination) and the PLSI-base approach, i.e., a single-feature-based approach. To investigate the effect on corpus size and quality for the clustering results, we apply the above clustering approaches to various types of text corpora such as Q and A documents, news articles, the Web and a balanced corpus. We also qualitatively analyze how CCRE extracts the other meanings of polysemous verbs.

After explaining the input data set in Sect. 4.1, we describe the evaluation method in Sect. 4.2 and the PLSI-based approach in Sect. 4.3, Finally we show the experimental results in Sect. 4.4.

Table 1 Statistics of input data: types of verb and noun pairs with case markers.

	pair	verb	nc	rt
Yahoo!	227985	8034	27405	1:3.41
M91	857688	11847	41433	1:3.50
M91-92	1530109	13381	56548	1:4.23
M91-93	2120935	14301	67833	1:4.74
M91-94	2745196	15246	80520	1:5.28
M91-95	3321608	15886	90305	1:5.68
M91-96	3921813	16391	100780	1:6.15
M91-97	4515098	16837	110550	1:6.57
M91-98	5085892	17204	119788	1:6.96
Web	12237162	13700	343119	1:25.05
BCCWJ	1879900	18023	91366	1:5.07
MSB	1438724	13219	54667	1:4.14

4.1 Input Data

The input data of the proposed clustering approach are the co-occurrence data between verbs and nouns with case markers. The co-occurrence data are constructed using a Japanese dependency parser from the following source text corpus.

(1) Yahoo! question and answer documents.
(2) Mainichi newspaper articles from 1991 to 1998.
(3) 500 million texts from the Web.
(4) Balanced Japanese corpus 2008 (BCCWJ) developed by The National Institute for Japanese Language.
(5) Mainichi newspaper articles which are the same size as BCCWJ[†] (MSB).

where Yahoo! Q&A contains spoken language expressions; MSB is prepared for measuring the effectivity of the genre variety of text corpora. Except for **(3)**, the texts were parsed by CaboCha a dependency parser[††]. **(3)** was parsed by KNP[†††].

Table 1 shows the statistics of the co-occurrence data of verb and noun-case pairs from each text corpus. In Table 1 'pair' denotes types of verbs and noun-cases, 'nc' denotes types of nouns with case markers, and 'rt' denotes the ratio of noun-cases to verbs. To see the effect of changing the amount of corpora, we prepared eight types of Mainichi newspaper article corpora varying from one to eight years. M91 denotes a one-year corpus of Mainichi newspaper articles in 1991, and M91-98 denotes eight years from 1991 to 1998. Table 1 shows that the Web corpus has a wider range of noun types to verbs than the other corpus because of the large amount of texts. BCCWJ also has a larger ratio of noun-cases to verbs than MSB as well as verb and noun-case types because it contains a wide genre of documents.

4.2 Evaluation Method

For evaluation of the synonymous verb groups extracted

[†]The amount is almost two years of articles.
[††]CaboCha: http://sourceforge.net/projects/cabocha/
[†††]KNP: http://www-lab25.kuee.kyoto-u.ac.jp/nl-resource/knp.html

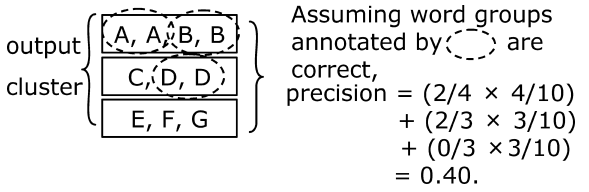

Fig. 3 Precision of output clusters.

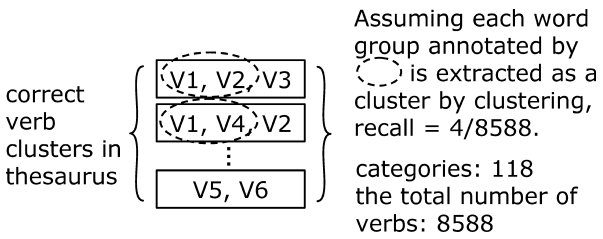

Fig. 4 Recall of thesaurus.

from the clustering system, we selected a free Japanese verb thesaurus manually constructed [12] as a gold standard. The thesaurus contains about 4400 verbs with five layers of the hierarchy. We calculated precision, which is sometimes called purity in clustering research [13], and recall on the basis of the verb categories at the 3rd hierarchy in the thesaurus[†]. Precision indicates how the output clusters correspond to the correct verb synonym groups, and recall indicates how the output clusters cover the correct categories in the verb thesaurus.

Examples of the calculation of precision and recall are shown in Figs. 3 and 4. As the basic idea of evaluating clusters we defined that a cluster should have a single correct group of verbs and noun-cases because the proposed approach is designed to extract a group of verbs and noun-cases with a shared meaning. If a cluster has several correct groups, we only regard the largest group in a cluster as the correct group. For example, in the first cluster in Fig. 3 only two elements were correctly extracted. This is the same in Fig. 4.

Even though the verb thesaurus is incomplete, the above precision and recall rates are useful to compare the performances of clustering approaches.

4.3 PLSI-Based Approach

We employed Hagiwara's synonym extraction approach [5] that is based on the probabilistic latent semantic indexing (PLSI) [14] method. Hagiwara's approach collects similar words by evaluating the similarity of two words on the basis of PLSI-based probability distribution.

Based on the PLSI formula, the probability that both the i-th verb and j-th noun-case will occur can be estimated by Eq. (4) using latent parameter z_k:

$$P(v_i, nc_j) = P(nc_j) \sum_{z_k} P(v_i|z_k)P(nc_j|z_k). \quad (4)$$

After $P(z|v, nc)$ is estimated using the EM algorithm, we can evaluate the similarity of two verbs (e.g., v_1 and v_2) by evaluating the distance between $P(z|v_1)$ and $P(z|v_2)$ by the Skew divergence[††]. This indicates that latent parameter z can be considered as decomposed feature vector.

The following are the steps of the synonym extraction approach: (1) set a key verb, (2) collect n verbs whose meanings are the most similar, (3) repeat steps (1) and (2) by shifting the key verb, and (4) delete duplicate clusters[†††]. This is not a clustering, but in each cluster similar verbs to the key verb are collected, and thus the precision of each cluster must be very high. By comparing the PLSI-based results with those of the proposed approach, we can see the effectivity of the co-clustering-based approach against the results of a single-feature-based approach.

One benefit of the single-feature-based approach is being able to take many features; however, in this paper, we only use the surface word co-occurrence data for the PLSI-based approach: the same input data of the co-clustering approach we proposed here. In the experimental results in Sect. 4.4 the number of latent semantics is set by z=200[††††], and the number of verbs for a cluster is set by n=5[†††††].

4.4 Experimental Results

We applied the PLSI-based approach, Aizawa's approach and the proposed CCRE to the input data set described in Sect. 4.1. Table 2 shows the precision[††††††], and Table 3 shows the recall. In the parentheses of Table 2 the numerator denotes the correctly extracted cluster elements, and the denominator denotes all of the output cluster elements registered in the thesaurus. In the parentheses of Table 3, the numerator denotes the correctly extracted cluster elements, and the denominator denotes all the cluster elements in the thesaurus.

In Table 2 the proposed CCRE shows the best performance in the number of extracted correct clusters of the

[†]At this hierarchy, antonyms are evaluated as the same category. The data are available athttp://vsearch.cl.cs.okayama-u.ac.jp/index.php

[††]The detailed setting is described in Hagiwara et al. [5].

[†††]The number of output clusters will be the same as the number of words, i.e., tokens. This is too many clusters when we compared the PLSI-based approach with the co-clustering-based approach in Tables 2 and 3. Thus we adjusted the number of PLSI's output clusters to 3000 because the output cluster number of the co-clustering approach was less than 2700.

[††††]We set this on the basis of preliminary experiment result.

[†††††]This number is defined to compare the same conditions of the co-clustering approach.

[††††††]In Table 2 we did not describe the correct cluster numbers because the number of correct clusters would be insufficient to evaluate the proposed approach's performance. This relates to how we count the correct elements in the output cluster in Fig. 3. The aim of our proposed approach is to support the human extraction of verb meanings. From the view of annotation task, it is not worth obtaining many clusters that only have a small number of correct elements; but obtaining a large number of correct elements from small clusters is worthwhile. Thus we did not use a cluster-number-based evaluation.

Table 2 Precision of PLSI, Aizawa's co-clustering, and CCRE.

	Precision		
Corpus	PLSI	Aizawa	CCRE
Yahoo!	0.118 (358/3043)	0.152 (507/3335)	0.169 (645/3812)
M91	0.158 (771/4883)	0.269 (1100/4083)	0.280 (2079/7421)
M91-92	0.188 (926/4935)	0.284 (1516/5331)	0.291 (2150/7378)
M91-93	0.195 (1045/5348)	0.289 (1650/5708)	0.298 (2444/8194)
M91-94	0.218 (1203/5508)	0.282 (1648/5854)	0.297 (2520/8471)
M91-95	0.203 (1084/5341)	0.315 (1788/5685)	0.325 (2678/8241)
M91-96	0.216 (1153/5337)	0.293 (1595/5452)	0.309 (2447/7926)
M91-97	0.229 (1174/5133)	0.307 (1631/5311)	0.317 (2465/7784)
M91-98	0.221 (1054/4779)	0.317 (1574/4967)	0.322 (2268/7037)
Web	0.164 (491/3000)	0.250 (1004/4020)	0.245 (1522/6221)
BCCWJ	0.153 (825/5379)	0.263 (1652/6293)	0.270 (2323/8618)
MSB	0.181 (874/4825)	0.279 (1485/5318)	0.292 (2126/7283)

Table 3 Recall of PLSI, Aizawa's co-clustering, and CCRE.

	Recall		
Corpus	PLSI	Aizawa	CCRE
Yahoo!	0.026 (225/8588)	0.047 (402/8588)	0.052 (448/8588)
M91	0.061 (525/8588)	0.078 (670/8588)	0.106 (914/8588)
M91-92	0.078 (673/8588)	0.103 (886/8588)	0.116 (994/8588)
M91-93	0.086 (738/8588)	0.113 (968/8588)	0.126 (1081/8588)
M91-94	0.092 (790/8588)	0.116 (995/8588)	0.131 (1128/8588)
M91-95	0.088 (757/8588)	0.130 (1113/8588)	0.148 (1270/8588)
M91-96	0.096 (821/8588)	0.122 (1046/8588)	0.138 (1185/8588)
M91-97	0.101 (866/8588)	0.117 (1009/8588)	0.135 (1160/8588)
M91-98	0.091 (783/8588)	0.118 (1010/8588)	0.133 (1146/8588)
Web	0.048 (409/8588)	0.086 (741/8588)	0.102 (872/8588)
BCCWJ	0.067 (572/8588)	0.128 (1097/8588)	0.145 (1241/8588)
MSB	0.076 (649/8588)	0.100 (860/8588)	0.113 (968/8588)

other approaches for all input corpora. Compared with Aizawa's approach, CCRE increases the number of correct clusters by about 50%[†].

CCRE's precision rate is higher than Aizawa's approach, and its average increase is 0.9%; on the other hand, the precision rate is only lower for the Web corpus for the following two reasons: the problem of the quality of the Web corpus and its size must be so large that the setting of the co-clustering approaches does not match the Web corpus. The same tendency can be seen in the M91-98 data. We discuss the second issue in Sect. 5. For the first case, a Web corpus has a lot of writing variations between Chinese characters (kanji) and the Japanese cursive syllabary (hiragana) for identical words. Even though the clustering approach can easily extract these writing variation verbs, the gold standard thesaurus does not contain the hiragana descriptions, decreasing the number of correct clusters. Clearly these writing variations are not needed, so this is not a thesaurus problem but instead reflects the quality of the Web corpus.

Table 3 shows that CCRE outperforms the other two approaches for all of the corpora in extracting verb clusters that match the thesaurus as well as the recall rates. Compared with Aizawa's approach, CCRE increases the number of correctly extracted verb elements in the thesaurus by about 15%[††]. The average increase of the recall rate is 1.5%.

As for the corpus size, both Aizawa's approach and CCRE show the best recall with the M91-95 corpus. The question remains: why is the M91-98 not the best for recall? Co-clustering apparently does not work well in large corpus. This will be discussed in Sect. 5.

The effectivity of a variety of genres can be seen by comparing precision between BCCWJ and MSB: the number of output correct clusters of BCCWJ is superior to that of MSB; however, BCCWJ's precision rate is worse than that of MSB. BCCWJ increases the recall rate about 30% against MSB, which is newspaper articles. The 30% increase corresponds to the improvement caused by adding three years of newspaper articles from M91-92 to M91-95 based on CCRE results in Table 3. This indicates that a balanced corpus can be an important factor to extract verb synonyms from text corpus.

5. Discussion

The scores of the precision and recall rates are not high due to the following two factors: the verb thesaurus is missing some verb meanings, and a problem exists when evaluating the recall rate of word meanings in a given corpus. To see the limitations of the above evaluation approach, we investigated how many verbs registered in the thesaurus failed to appear in each corpus and estimate how many verb meanings exist to be extracted in Table 4.

Vt denotes the verb types that appeared in both the cor-

[†]The total number of correct cluster elements in Aizawa's approach for all corpora is 15502, and the total in CCRE is 23344. The increasing rate by CCRE compared with Aizawa's approach is 23344/15502 = 1.506; and thus CCRE boosts the numbers of correctly extracted cluster elements about 50% more than Aizawa's approach.

[††]The total number of correctly extracted verbs in the thesaurus with Aizawa's approach for all corpora is 10797, and the total by CCRE is 12407. Compared with Aizawa's approach, the rate of increase by CCRE is 12407/10797 = 1.149; and thus CCRE boosts the numbers of correctly extracted verbs about 15% more than Aizawa's approach.

Table 4 Estimation of verb meanings in a corpus.

Corpus	Vt	Evmt	Vmt
Yahoo!	1744	3774	8588
M91	3572	6195	8588
M91-92	3905	6670	8588
M91-93	4011	6815	8588
M91-94	4119	7008	8588
M91-95	4186	7103	8588
M91-96	4247	7196	8588
M91-97	4302	7285	8588
M91-98	4393	7353	8588
Web	3000	4947	8588
BCCWJ	3879	6899	8588
MSB	3886	6644	8588

pus and the thesaurus. Evmt denotes the estimated number of verb meanings that appear in both the thesaurus and the targeted corpus. We estimated the numbers by counting all the meanings of verbs in the thesaurus when they appear in the corpus without noting their meanings in the corpus[†]. Evmt indicates the numbers of possible verb meanings that can be extracted from the corpus. Vmt denotes all the verb meanings in the thesaurus.

Table 4 reveals that the verb types in the Web corpus are very different from those in the thesaurus. This is a reason why the performance of the Web corpus is not high. For the other corpora, the verb meanings (types) appear with an increase of size. Let us point out again the Evmts in Table 4 are larger than the real numbers of verb meanings that are supposed to be extracted. The maximum number of extractable verb meanings in the thesaurus can be much smaller than the Evmts in Table 4 since the variety of verb meanings in news paper articles are usually limited[††]. To know the real numbers of Evmts we have to manually analyze all of the verb meanings in each corpus, which is basically impossible.

We have not executed manual evaluations of the output clusters because the trend of the above results suggests that the best CCRE performance is the same as our intuition when we checked a very small number of output clusters. Even though we currently see a precision rate in CCRE of about 60%, we still need to do accurate evaluations.

In the above experimental results, the largest corpus such as M91-98 does not provide the best clustering performance. One reason might be the fixed link limitation of the co-clustering approach. The aim of the limitation is to filter out words that have too many links such as "suru" (do) or "aru" (is). Thus the co-cluster approach deletes words whose link variations exceed 100 link types. If we allow a more tolerant link number with a bigger corpus, accuracy will be improved. Table 5 shows the precision and recall varying link sizes on M91-98.

From Table 5 all of the precision rates with link sizes from 110 to 300 exceed those with a link size of 100. The best precision rate for M91-98 is 0.329 with a link size of 110 or 125. This score is higher than the best precision rate i.e., 0.325, which is shown in Table 4 when applying CCRE to the M91-95 corpus with a link size of 100. This indicates

Table 5 Precision and Recall varying link size on M91-98 by CCRE.

Link size	Precision	Recall
100	0.322(2268/7037)	0.133(1146/8588)
110	0.329(2838/8616)	0.151(1297/8588)
125	0.329(3410/10362)	0.166(1429/8588)
150	0.327(4352/13312)	0.189(1619/8588)
200	0.326(6197/19037)	0.223(1917/8588)
300	0.325(9420/29014)	0.264(2271/8588)

Table 6 Example where CCRE extracts different meanings of polysemous verbs.

(a) haneagaru (become expensive/rise)		
1st	verb	kousyou (become expensive), sagaru (drop), nesagari (fall in price), neagari (become expensive), jousyou (boost), haneagaru (become expensive)
	noun	en-ni (yen), kakaku-ga (price), bukka-ga (price), bai-ni (double)
2nd	verb	takanaru (pulse), odoru (pulse), chijimu (be surprised), yowaru (become weak), haneagaru (rise)
	noun	handou-de (reaction), shintai-ga (body), shinzou-ga (heart), kun-to (onomatopoeia)
(b) takuwaeru (grow a beard/save)		
1st	verb	takuwaeru (grow a beard/save), takuwaeru (grow a beard/save), suiageru (suck up), hayasu (wear), hayasu (wear)
	noun	hige-wo (beard), kuchihige-wo (beard), mitsu-wo (honey), eiyou-wo (nutrishment), seiryoku-wo (power)
2nd	verb	takuwaeru (save), hokyuu (supply), chikuseki (store), umidasu (generate) umu (generate)
	noun	rishi-wo (interest), zaigen-wo (resource), kaikei-kara (accounting), zouzei-de (tax increase), kokusai-de (government bonds), kaikei-de (accounting)

that CCRE might obtain higher precision rates for larger corpora. The recall rate in Table 5 also increases linearly with link size. When the link size is 300 the recall rate (0.264) is about twice as high as the recall rate (0.133) when the link size is 100. Currently CCRE requires explosive computation time for large link size: for example, in the case of a link size of 300, seven days are needed with Xeon 5150 in Dell Precision 690. However, since the CCRE algorithm can be parallelized, if we can employ many CPUs this will not be a serious problem.

To see CCRE's effectivity for extracting other meanings of a polysemous verb, Table 6 shows the cluster examples that were recursively extracted by CCRE. The extracted clusters of the two verbs are "haneagaru" (become expensive/rise) and "takuwaeru" (grow a beard/save)[†††]. The '1st' and '2nd' in the left column in the table denote the extrac-

[†]Verbs that appear in the corpus less than two times are not counted.

[††]For example, we found that Japanese verb nyuuin-suru (be hospitalized/become a Buddhist monk) is only used as the meaning of "be hospitalized" in the news corpus.

[†††]In Table 6 the words such as "takuwaeru" and "hayasu" exist doubly at the '1st' cluster of **(b)** because of writing variation between Chinese characters and the Japanese cursive syllabary.

tion time of each cluster: '1st' indicates that the cluster was extracted from initial bipartite graph that is the same as Aizawa's approach; and '2nd' indicates that the cluster was extracted from the eliminated initial bipartite graph. If the meaning of the '2nd' verb cluster is different from the '1st' one, CCRE successfully extracted a different meaning.

For "haneagaru" the meaning of the '1st' verb cluster is *become expensive*, and the '2nd' cluster is *rise*. Thus in this case CCRE successfully extracted a different meaning. The meaning of the '1st' cluster of "takuwaeru" is *grow a beard*, but the cluster seems a little noisy because it contains an unrelated verb "suiageru" (suck up), and noun-cases "mitsu-wo" (honey) and "eiyou-wo" (nourishment). The meanings of the '2nd' cluster are *save* as well as unrelated meaning, *generate*. From view of clustering accuracy, the quality of the cluster is not high; but from the view of an tool aimed to construct a verb synonym dictionary, because a human annotator can recognize that two such verbs as "takuwaeru" and "chikuseki" compose a cluster whose meaning is *save*. These results show that our proposed CCRE clustering approach is promising.

6. Conclusion

In this paper we proposed co-clustering with recursive elimination to extract the different meanings of polysemous verbs. From the experimental results of verb synonym extraction, our proposed approach outperformed Aizawa's co-clustering approach and a PLSI-based synonym extraction approach. The following are the characteristics of our proposed approach: (1) it can deal with a large-scale input text corpus, (2) it can simultaneously extract two elements such as verbs and nouns, and (3) it has a function to actively extract other meanings of polysemous words. We believe that the proposed co-clustering approach is highly promising.

Acknowledgement

This research was supported by a Grant-in-Aid for Scientific Research on Priority Areas, "Japanese Corpus" from the Ministry of Education, Culture, Sports, Science and Technology of Japan.

References

[1] D. Hindle, "Noun classification from predicate-argument structures," Proc. 28th Annual Meeting of The Association for Computational Linguistics, pp.268–275, 1990.
[2] F. Pereira and N. Tishby, "Distributional clustering of English words," Proc. 31st Annual Meeting of The Association for Computational Linguistics, pp.183–190, 1993.
[3] M. Sasaki and H. Shinnou, "Automatic thesaurus construction using word clustering," Proc. Pacific Association for Computational Linguistics 2003, pp.55–62, 2003.
[4] D. Mochihashi and Y. Matsumoto, "Probabilistic representation of meaning," IPSJ SIG NL Technical Reports, vol.4, pp.77–84, 2002.
[5] M. Hagiwara, Y. Ogawa, and K. Toyama, "Utilization of plsi for automated thesaurus construction," IPSJ SIG Technical Reports, 2005-NL-166, pp.71–78, 2005.
[6] C. Kemp, J.B. Tenenbaum, T.L. Griffiths, T. Yamada, and N. Ueda, "Learning systems of concepts with an infinite relational model," Proc. 21st National Conference on Artificial Intelligence, pp.381–388, 2006.
[7] K. Kurihara, Y. Kameya, and T. Sato, "Discovering concepts from word co-occurrences with a relational model," Japanese Society of Artificial Intelligence, vol.22, no.2, pp.218–226, 2007.
[8] I.S. Dhillon, "Co-clustering documents and using bipartite spectral graph partitioning," Proc. 7th ACM SIGKDD, pp.269–274, 2001.
[9] I.S. Dhillon, S. Mallela, and D.S. Modha, "Information-theoretic co-clustering," Proc. 9th ACM SIGKDD, pp.89–98, 2003.
[10] K. Takeuchi, "Extraction of verb synonyms using co-clustering approach," Proc. 2nd International Symposium on Universal Communication, pp.173–178, Osaka, Japan, Dec. 2008.
[11] A. Aizawa, "A method of cluster-based indexing of textual data," Proc. COLING 2002, pp.1–7, 2002.
[12] K. Takeuchi, K. Inui, N. Takeuchi, and A. Fujita, "Detailed categories of verb argument structure on the basis of semantic," Proc. 14th Annual Meeting of The Association for Natural Language Processing, pp.1037–1040, 2008.
[13] A. Hotho, S. Staab, and G. Stumme, "Wordnet improves text document clustering," Proc. SIGIR 2003 Semantic Web Workshop, 2003.
[14] T. Hofmann, "Probabilistic latent semantic indexing," Proc. 22nd International Conference on Research and Development in Informatin Retrieval (SIGIR 99), pp.50–57, 1999.

Koichi Takeuchi received a Doctor of Engineering from the Nara Institute of Science and Technology, Japan in 1998. He is currently a Senior Assistant Professor of the Graduate School of Natural Science and Technology, Okayama University, Japan. His research interests include the construction of language understanding systems.

Hideyuki Takahashi received a B.E. from Okayama University in Japan in 2009. He is currently a masters student in the Graduate School of Natural Science and Technology, Okayama University, Japan. His research interests include the word sense disambiguation.