Proceedings of HLT-NAACL 2003
Main Papers , pp. 32-39
Edmonton, May-June 2003

Word Sense Acquisition from Bilingual Comparable Corpora

Hiroyuki Kaji

Central Research Laboratory, Hitachi, Ltd.
1-280 Higashi-Koigakubo, Kokubunji-shi, Tokyo 185-8601, Japan
kaji@crl.hitachi.co.jp

Abstract

Manually constructing an inventory of word senses has suffered from problems including high cost, arbitrary assignment of meaning to words, and mismatch to domains. To overcome these problems, we propose a method to assign word meaning from a bilingual comparable corpus and a bilingual dictionary. It clusters second-language translation equivalents of a first-language target word on the basis of their translingually aligned distribution patterns. Thus it produces a hierarchy of corpus-relevant meanings of the target word, each of which is defined with a set of translation equivalents. The effectiveness of the method has been demonstrated through an experiment using a comparable corpus consisting of Wall Street Journal and Nihon Keizai Shimbun corpora together with the EDR bilingual dictionary.

1 Introduction

Word Sense Disambiguation (WSD) is an important subtask that is necessary for accomplishing most natural language processing tasks including machine translation and information retrieval. A great deal of research on WSD has been done over the past decade (Ide and Veronis, 1998). In contrast, word sense acquisition has been a human activity; inventories of word senses have been constructed by lexicographers based on their intuition. Manually constructing an inventory of word senses has suffered from problems such as high cost, arbitrary division of word senses, and mismatch to application domains.

We address the problem of word sense acquisition along the lines of the WSD where word senses are defined with sets of translation equivalents in another language. Bilingual corpora or second-language corpora enable unsupervised WSD (Brown, et al., 1991; Dagan and Itai, 1994). However, the correspondence between senses of a word and its translations is not one-to-one, and therefore we need to prepare an inventory of word senses, each of which is defined with a set of synonymous translation equivalents. Although conventional bilingual dictionaries usually group translations according to their senses, the grouping differs by dictionary. In addition, senses specific to a domain are often missing while many senses irrelevant to the domain or rare senses are included. To overcome these problems, we propose a method for producing a hierarchy of clusters of translation equivalents from a bilingual corpus and a bilingual dictionary.

To the best of our knowledge, there are two preceding research papers on word sense acquisition (Fukumoto and Tsujii, 1994; Pantel and Lin, 2002). Both proposed distributional word clustering algorithms that are characterized by their capabilities to produce overlapping clusters. According to their algorithms, a polysemous word is assigned to multiple clusters, each of which represents one of its senses. These and our approach differ in how to define the word sense, i.e., a set of synonyms in the same language versus a set of translation equivalents in another language. Schuetze (1998) proposed a method for dividing occurrences of a word into classes, each of which consists of contextually similar occurrences. However, it does not produce definitions of senses such as sets of synonyms and sets of translation equivalents.

2 Basic Idea

2.1 Clustering of translation equivalents

Most work on automatic extraction of synonyms from text corpora rests on the idea that synonyms have

similar distribution patterns (Hindle, 1990; Peraira, et al., 1993; Grefenstette, 1994). This idea is also useful for our task, i.e., extracting sets of synonymous translation equivalents, and we adopt the approach to distributional word clustering.

We need to mention that the singularity of our task makes the problem easier. First, we do not have to cluster all words of a language, but we only have to cluster a small number of translation equivalents for each target word, whose senses are to be extracted, separately. As a result, the problem of computational efficiency becomes less serious. Second, even if a translation equivalent itself is polysemous, it is not necessary to consider senses that are irrelevant to the target word. A translation equivalent usually represents one and only one sense of the target word, at least in case the language-pair is those with different origins like English and Japanese. Therefore, a non-overlapping clustering algorithm, which is far simpler than overlapping clustering algorithms, is sufficient.

2.2 Translingual distributional word clustering

In conventional distributional word clustering, a word is characterized by a vector or weighted set consisting of words in the same language as that of the word itself. In contrast, we propose a translingual distributional word clustering method, whereby a word is characterized by a vector or weighted set consisting of words in another language. It is based on the sense-vs.-clue correlation matrix calculation method we originally developed for unsupervised WSD (Kaji and Morimoto, 2002). That method presupposes that each sense of a target word x is defined with a synonym set consisting of the target word itself and one or more translation equivalents which represent the sense. It calculates correlations between the senses of x and the words statistically related to x, which act as clues for determining the sense of x, on the basis of translingual alignment of pairs of related words. Rows of the resultant correlation matrix are regarded as translingual distribution patterns characterizing translation equivalents.

Sense-vs.-clue correlation matrix calculation method *)

1) Alignment of pairs of related words

Let $X(x)$ be the set of clues for determining the sense of a first-language target word x. That is,

$X(x)=\{x' \mid (x, x') \in R_X\}$,

where R_X denotes the collection of pairs of related words extracted from a corpus of the first language. Henceforth, the j-th clue for determining the sense of x will be denoted as $x'(j)$. Furthermore, let $Y(x, x'(j))$ be the set consisting of all second-language counterparts of a first-language pair of related words x and $x'(j)$. That is,

$Y(x, x'(j)) = \{(y, y') \mid (y, y') \in R_Y, (x, y) \in D, (x'(j), y') \in D\}$,

where R_Y denotes the collection of pairs of related words extracted from a corpus of the second language, and D denotes a bilingual dictionary, i.e., a collection of pairs consisting of a first-language word and a second-language word that are translations of one another.

Then, for each alignment, i.e., pair of $(x, x'(j))$ and (y, y') $(\in Y(x, x'(j)))$, a weighted set of common related words $Z((x, x'(j)), (y, y'))$ is constructed as follows:

$Z((x, x'(j)), (y, y')) = \{x'' / w(x'') \mid (x, x'') \in R_X, (x'(j), x'') \in R_X\}$.

The weight of x'', denoted as $w(x'')$, is determined as follows:

- $w(x'') = 1+\alpha \cdot MI(y, y')$ when $\exists y''$ $(x'', y'') \in D$, $(y, y'') \in R_Y$, and $(y', y'') \in R_Y$.
- $w(x'') = 1$ otherwise.

This is where $MI(y, y')$ is the mutual information of y and y'. The coefficient α was set to 5 in the experiment described in Section 4.

2) Calculation of correlation between senses and clues

The correlation between the i-th sense $S(x, i)$ and the j-th clue $x'(j)$ is defined as:

$$C(S(x,i), x'(j)) = MI(x, x'(j)) \cdot \frac{\max_{\substack{(y,y') \in Y(x,x'(j)), \\ y \in S(x,i)}} A((x, x'(j)), (y, y'), S(x,i))}{\max_k \left\{ \max_{\substack{(y,y') \in Y(x,x'(j)), \\ y \in S(x,k)}} A((x, x'(j)), (y, y'), S(x,k)) \right\}}.$$

This is where $MI(x, x'(j))$ is the mutual information of x and $x'(j)$, and $A((x, x'(j)), (y, y'), S(x,i))$, the plausibility of alignment of $(x, x'(j))$ with (y, y') suggesting $S(x, i)$, is defined as the weighted sum of the correlations between the sense and the common related words, i.e.,

$$A((x, x'(j)), (y, y'), S(x,i)) = \sum_{x'' \in Z((x,x'(j)),(y,y'))} w(x'') \cdot C(S(x,i), x'').$$

The correlations between senses and clues are cal-

*) A description of the wild-card pair of related words, which plays an essential role in recovering alignment failure, has been omitted for simplicity.

culated iteratively with the following initial values: $C_0(S(x, i), x'(j))=MI(x, x'(j))$. The number of iterations was set to 6 in the experiment. Figure 1 shows how the correlation values converge.

Advantages of using translingually aligned distribution patterns

Translingual distributional word clustering has advantages over conventional monolingual distributional word clustering, when they are used to cluster translation equivalents of a target word. First, it avoids clusters being degraded by polysemous translation equivalents. Let "race" be the target word. One of its translation equivalents, "レース<*REESU*>", is a polysemous word representing "lace" as well as "race". According to monolingual distributional word clustering, "レース<*REESU*>" is characterized by a mixture of the distribution pattern for "レース<*REESU*>" representing "race" and that for "レース<*REESU*>" representing "lace", which often results in degraded clusters. In contrast, according to translingual distributional word clustering, "レース<*REESU*>" is characterized by the distribution pattern for the sense of "race" that means "competition".

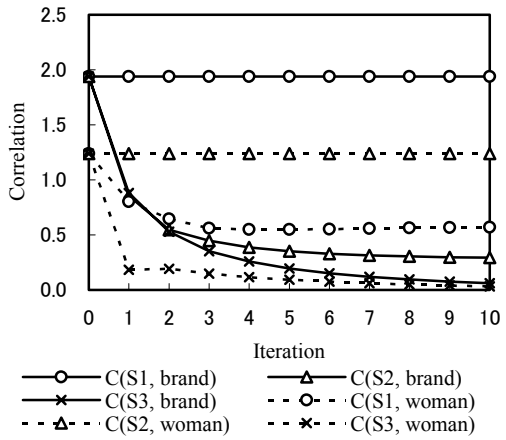
Second, translingual distributional word clustering can exclude from the clusters translation equivalents irrelevant to the corpus. For example, a bilingual dictionary renders "特徴<*TOKUCHOU*>" ("feature") as a translation of "race", but that sense of "race" is used infrequently. If it is the case in a given domain, "特徴<*TOKUCHOU*>" has low correlation with most words related to "race", and can therefore be excluded from any clusters.

We should also mention the data-sparseness problem that hampers distributional word clustering. Generally speaking, the problem becomes more difficult in translingual distributional word clustering, since the sparseness of data in two languages is multiplied. However, the sense-vs.-clue correlation matrix calculation method overcomes this difficulty; it calculates the correlations between senses and clues iteratively to smooth out the sparse data.

Translingual distributional word clustering can also be implemented on the basis of word-for-word alignment of a parallel corpus. However, availability of large parallel corpora is extremely limited. In contrast, the sense-vs.-clue correlation calculation method accepts comparable corpora which are available in many domains.

2.3 Similarity based on subordinate distribution pattern

Naive translingual distributional word clustering based



S1={promotion, 宣伝<*SENDEN*>, プロモーション<*PUROMOUSHON*>, 売り込み<*URIKOMI*>, …}
("an activity intended to help sell a product")
S2={promotion, 昇格<*SHOUKAKU*>, 昇進<*SHOUSHIN*>, 登用<*TOUYOU*>, …}
("advancement in rank or position")
S3={promotion, 奨励<*SHOUREI*>, 振興<*SHINKOU*>, 助長<*JOCHOU*>,…}
("action to help something develop or succeed")

Figure 1. Convergence of correlation between senses and clues.

on the sense-vs.-clue correlation matrix calculation method is outlined in the following steps:

1) Define the sense of a target word by using each translation equivalent.
2) Calculate the sense-vs.-clue correlation matrix for the set of senses resulting from step 1).
3) Calculate similarities between senses on the basis of distribution patterns shown by the sense-vs.-clue correlation matrix.
4) Cluster senses by using a hierarchical agglomerative clustering method, e.g., the group-average method.

However, this naive method is not effective because some senses usually have duplicated definitions in step 1) despite the fact that the sense-vs.-clue correlation matrix calculation algorithm presupposes a set of senses without duplicated definitions. The algorithm is based on the "one sense per collocation" hypothesis, and it results in each clue having a high correlation with one and only one sense. A clue can never have high correlations with two or more senses, even when they are actually the same sense. Consequently, synonymous translation equivalents do not necessarily have high similarity.

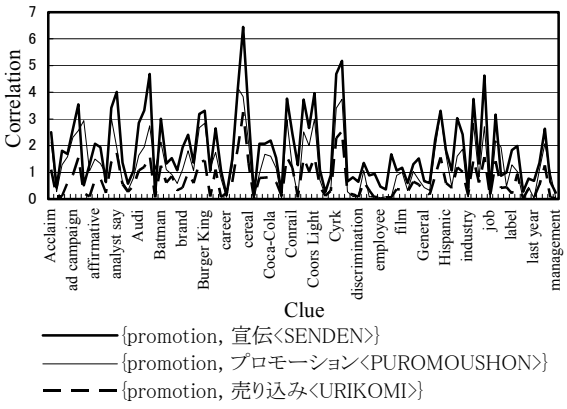
Figure 2(a) shows parts of distribution patterns for

{promotion, 宣伝<*SENDEN*>}, {promotion, プロモーション<*PUROMOUSHON*>}, and {promotion, 売り込み<*URIKOMI*>} all of which define the "sales activity" sense of "promotion". We see that most clues for selecting that sense have higher correlation with {promotion, 宣伝<*SENDEN*>} than with {promotion, プロモーション<*PUROMOUSHON*>} and {promotion, 売り込み<*URIKOMI*>}. This is because "宣伝<*SENDEN*>" is the most dominant translation equivalent of "promotion" in the corpus.

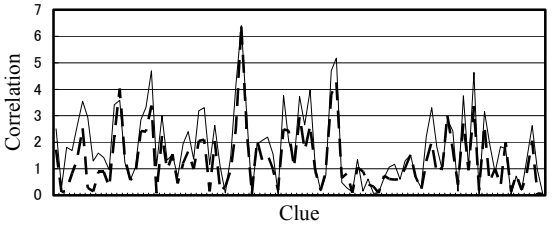
To resolve the above problem, we calculated the sense-vs.-clue correlation matrix not only for the full set of senses but also for the set of senses excluding one of these senses. Excluding a definition of the sense, which includes the most dominant translation equivalent, allows most clues for selecting the sense to have the highest correlations with another definition of the same sense, which includes the second most dominant translation equivalent. Figure 2(b) shows parts of distribution patterns for {promotion, プロモーション<*PUROMOUSHON*>} and {promotion, 売り込み<*URIKOMI*>} shown by the sense-vs.-clue correlation matrix for the set of senses excluding {promotion, 宣伝<*SENDEN*>}. We see that most clues for selecting the "sales activity" sense have higher correlations with {promotion, プロモーション<*PUROMOUSHON*>} than with {promotion, 売り込み<*URIKOMI*>}. This is because "プロモーション<*PUROMOUSHON*>" is the second most dominant translation equivalent in the corpus. We also see that the distribution pattern for {promotion, プロモーション<*PUROMOUSHON*>} in Fig. 2(b) is more similar to that for {promotion, 宣伝<*SENDEN*>} in Fig. 2(a) than that for {promotion, プロモーション<*PUROMOUSHON*>} in Fig. 2(a).

We call the distribution pattern for sense S_2, resulting from the sense-vs.-clue correlation matrix for the set of senses excluding sense S_1, the distribution pattern for S_2 subordinate to S_1, while we call the distribution pattern for sense S_2, resulting from the sense-vs.-clue correlation matrix for the full set of senses, simply the distribution pattern for S_2. We define the similarity of S_2 to S_1 as the similarity of the distribution pattern for S_2 subordinate to S_1 to the distribution pattern for S_1.

Calculating the sense-vs.-clue correlation matrix for a set of senses excluding one sense is of course insufficient since three or more translation equivalents may represent the same sense of the target word. We should calculate the sense-vs.-clue correlation matrices both for the full set of senses and for the set of senses excluding one of these senses again, after merging similar senses into one. Repeating these procedures enables corpus-relevant but less dominant translation equivalents to be drawn up, while corpus-irrelevant ones are never drawn up. Thus, a hierarchy of corpus-relevant senses or clusters of corpus-relevant translation equivalents is produced.



(a) Distribution patterns



(b) Distribution patterns subordinate to {promotion, 宣伝<*SENDEN*>}

Figure 2. Distribution Patterns for Some Senses of "promotion".

3 Proposed Method

3.1 Outline

As shown in Fig. 3, our method repeats the following three steps:

1) Calculate sense-vs.-clue correlation matrices both for the full set of senses and for a set of senses excluding each of these senses.
2) Calculate similarities between senses on the basis of distribution patterns and subordinate distribution patterns.
3) Merge each pair of senses with high similarity into one.

The initial set of senses is given as $\Sigma(x)=\{\{x, y_1\}, \{x, y_2\}, \ldots, \{x, y_N\}\}$ where x is a target word in the first language, and y_1, y_2, …, and y_N are translation equivalents of x in the second-language. Translation equivalents that occur less frequently in the second-language corpus can be excluded from the initial

set to shorten the processing time. The details of the steps are described in the following sections.

3.2 Calculation of sense-vs.-clue correlation matrices

First, a sense-vs.-clue correlation matrix is calculated for the full set of senses. The resulting correlation matrix is denoted as C. That is, $C(i, j)$ is the correlation between the i-th sense $S(x,i)$ of a target word x and its j-th clue $x'(j)$.

Then a set of active senses, $\Sigma_A(x)$, is determined. A sense is regarded active if and only if the ratio of clues with which it has the highest correlation exceeds a predetermined threshold θ (In the experiment in Section 4, θ was set to 0.05). That is,

$$\Sigma_A(x) = \{S(x,i) \mid R(S(x,i)) > \theta\},$$

where $R(S(x, i))$ denotes the ratio of clues having the highest correlation with $S(x, i)$, i.e.,

$$R(S(x,i)) = \left|\{x'(j) \mid C(i,j) = \max_k C(k,j)\}\right| \Big/ \left|\{x'(j)\}\right|.$$

Thus $\Sigma_A(x)$ consists of senses of the target word x that are relevant to the corpus.

Finally, a sense-vs.-clue correlation matrix is calculated for the set of senses excluding each of the active senses. The correlation matrix calculated for the set of senses excluding the k-th sense is denoted as C_{-k}. That is, $C_{-k}(i, j)$ ($i \neq k$) is the correlation between the i-th sense and the j-th clue that is calculated excluding the k-th sense. $C_{-k}(k, j)$ (j=1, 2, ...) are set to zero. This redundant k-th row is included to maintain the same correspondence between rows and senses as in C.

3.3 Calculation of sense similarity matrix

Similarity of the i-th sense $S(x, i)$ to the j-th sense $S(x, j)$, $Sim(S(x, i), S(x, j))$, is defined as the similarity of the distribution pattern for $S(x, i)$ subordinate to $S(x, j)$ to the distribution pattern of $S(x, j)$. Note that this similarity is asymmetric and reflects which sense is more dominant in the corpus. It is probable that $Sim(S(x, i), S(x, j))$ is large but $Sim(S(x, j), S(x, i))$ is not when $S(x, j)$ is more dominant than $S(x, i)$.

According to the sense-vs.-clue correlation matrix, each sense is characterized by a weighted set of clues. Therefore, we used the weighted Jaccard coefficient as the similarity measure. That is,

$$Sim(S(x,i),S(x,j)) = \frac{\sum_k \min\{C_{-j}(i,k), C(j,k)\}}{\sum_k \max\{C_{-j}(i,k), C(j,k)\}}$$

when $S(x, j) \in \Sigma_A(x)$.

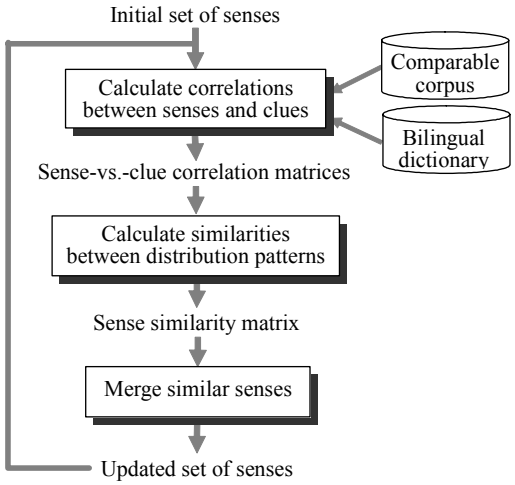


Figure 3. Flow Diagram of Proposed Method.

$Sim(S(x,i),S(x,j)) = 0$ otherwise.

It should be noted that a sense is characterized by different weighted sets of clues depending on which sense the similarity is calculated. Note also that inactive senses are neglected because they are not reliable.

3.4 Merging similar senses

The set of senses is updated by merging every pair of mutually most-similar senses into one. That is,

$$\Sigma(x) \leftarrow \Sigma(x) - \{S(x, i), S(x, j)\} + \{S(x, i) \cup S(x, j)\}$$

if $Sim(S(x,i),S(x,j)) = \max_{j'}\{\max$

$\{Sim(S(x,i),S(x,j')),Sim(S(x,j'),S(x,i))\}\}$,

$Sim(S(x,i),S(x,j)) = \max_{i'}\{\max$

$\{Sim(S(x,i'),S(x,j)),Sim(S(x,j),S(x,i'))\}\}$,

and $Sim(S(x,i),S(x,j)) > \sigma$.

The σ is a predetermined threshold for similarity, which is introduced to avoid noisy pairs of senses being merged. In the experiment in Section 4, σ was set to 0.25.

If at least one pair of senses are merged, the whole procedure, i.e., the calculation of sense-vs.-clue matrices through the merger of similar senses, is repeated for the updated set of senses. Otherwise, the clustering procedure terminates.

Agglomerative clustering methods usually suffer from the problem of when to terminate merging. In our method described above, the similarity of senses that are merged into one does not necessarily decrease

monotonically, which makes the problem more difficult. At present, we are forced to output a dendrogram that represents the history of mergers and leave the final decision to humans. The dendrogram consists of translation equivalents that are included in active senses in the final cycle. Other translation equivalents are rejected as they are irrelevant to the corpus.

4 Experimental Evaluation

4.1 Experimental settings

Our method was evaluated through an experiment using a Wall Street Journal corpus (189 Mbytes) and a Nihon Keizai Shimbun corpus (275 Mbytes).

First, collected pairs of related words, which we restricted to nouns and unknown words, were obtained from each corpus by extracting pairs of words co-occurring in a window, calculating mutual information of each pair of words, and selecting pairs with mutual information larger than the threshold. The size of the window was 25 words excluding function words, and the threshold for mutual information was set to zero. Second, a bilingual dictionary was prepared by collecting pairs of nouns that were translations of one another from the Japan Electronic Dictionary Research Institute (EDR) English-to-Japanese and Japanese-to-English dictionaries. The resulting dictionary includes 633,000 pairs of 269,000 English nouns and 276,000 Japanese nouns.

Evaluating the performance of word sense acquisition methods is not a trivial task. First, we do not have a gold-standard sense inventory. Even if we have one, we have difficulty mapping acquired senses onto those in it. Second, there is no way to establish the complete set of senses appearing in a large corpus. Therefore, we evaluated our method on a limited number of target words as follows.

We prepared a standard sense inventory by selecting 60 English target words and defining an average of 3.4 senses per target word manually. The senses were rather coarse-grained; i.e., they nearly corresponded to groups of translation equivalents within the entries of everyday English-Japanese dictionaries. We then sampled 100 instances per target word from the Wall Street Journal corpus, and we sense-tagged them manually. Thus, we estimated the ratios of the senses in the training corpus for each target word.

We defined two evaluative measures, recall of senses and accuracy of sense definitions. The recall of senses is the proportion of senses with ratios not less than a threshold that are successfully extracted, and it varies with change of the threshold. We judged that a sense was extracted, when it shared at least one translation equivalent with some active sense in the final cycle.

To evaluate the accuracy of sense definitions while avoiding mapping acquired senses onto those in the standard sense inventory, we regard a set of senses as a set of pairs of synonymous translation equivalents. Let T_S be a set consisting of pairs of translation equivalents belonging to the same sense in the standard sense inventory. Likewise, let $T(k)$ be a set consisting of pairs of translation equivalents belonging to the same active sense in the k-th cycle. Further, let U be a set of pairs of translation equivalents that are included in active senses in the final cycle. Recall and precision of pairs of synonymous translation equivalents in the k-th cycle are defined as:

$$R(k) = \frac{|T_S \cap T(k)|}{|T_S \cap U|}.$$

$$P(k) = \frac{|T_S \cap T(k)|}{|T(k)|}.$$

Further, F-measure of pairs of synonymous translation equivalents in the k-th cycle is defined as:

$$F(k) = \frac{2 \cdot R(k) \cdot P(k)}{R(k) + P(k)}.$$

The F-measure indicates how well the set of active senses coincides with the set of sense definitions in the standard senses inventory. Although the current method cannot determine the optimum cycle, humans can identify the set of appropriate senses from a hierarchy of senses at a glance. Therefore, we define the accuracy of sense definitions as the maximum F-measure in all cycles.

4.2 Experimental results

To simplify the evaluation procedure, we clustered translation equivalents that were used to define the senses of each target word in the standard sense inventory, rather than clustering translation equivalents rendered by the EDR bilingual dictionary. The recall of senses for totally 201 senses of the 60 target words was:

96% for senses with ratios not less than 25%,

87% for senses with ratios not less than 5%, and

78% for senses with ratios not less than 1%.

The accuracy of sense definitions, averaged over the 60 target words, was 77%.

The computational efficiency of our method proved to be acceptable. It took 13 minutes per target word on a HP9000 C200 workstation (CPU clock: 200 MHz, memory: 32 MB) to produce a hierarchy of clusters of translation equivalents.

Some clustering results are shown in Fig. 4. These demonstrate that our proposed method shows a great deal of promise. At the same time, evaluating the results revealed its deficiencies. The first of these lies in the crucial role of the bilingual dictionary. It is obvious that a sense is never extracted if the translation equivalents representing it are not included in it. An exhaustive bilingual dictionary is therefore required. From this point of view, the EDR bilingual dictionary is fairly good. The second deficiency lies in the fact that it performs badly for low-frequency or non-topical senses. For example, the sense of "bar" as the "legal profession" was clearly extracted, but its sense as a "piece of solid material" was not extracted.

We also compared our method with two alternatives: monolingual distributional clustering mentioned in Section 2.2 and naive translingual clustering mentioned in Section 2.3. Figures 5(a), (b), and (c) show respective examples of clustering obtained by our method, the monolingual method, and the naive translingual method. Comparing (a) with (b) reveals the superiority of the translingual approach to the monolingual approach, and comparing (a) with (c) reveals the effectiveness of the subordinate distribution pattern introduced in Section 2.3. Note that deleting the corpus-irrelevant translation equivalents from the dendrograms in both (b) and (c) would not result in appropriate ones.

5 Discussion

Our method has several practical advantages. One of these is that it produces a corpus-dependent inventory of word senses. That is, the resulting inventory covers most senses relevant to a domain, while it excludes senses irrelevant to the domain.

Second, our method unifies word sense acquisition with word sense disambiguation. The sense-vs.-clue correlation matrix is originally used for word sense disambiguation. Therefore, our method guarantees that acquired senses can be distinguished by machines, and further it demonstrates the possibility of automatically optimizing the granularity of word senses.

Some limitations of the present methods are discussed in the following with possible future extensions. First, our method produces a hierarchy of clusters but cannot produce a set of disjoint clusters. It is very important to terminate merging senses autonomously during an appropriate cycle. Comparing distribution patterns (not subordinate ones) may be useful to terminate merging; senses characterized by complementary distribution patterns should not be merged.

Second, the present method assumes that each translation equivalent represents one and only one sense of the target word, but this is not always the case.

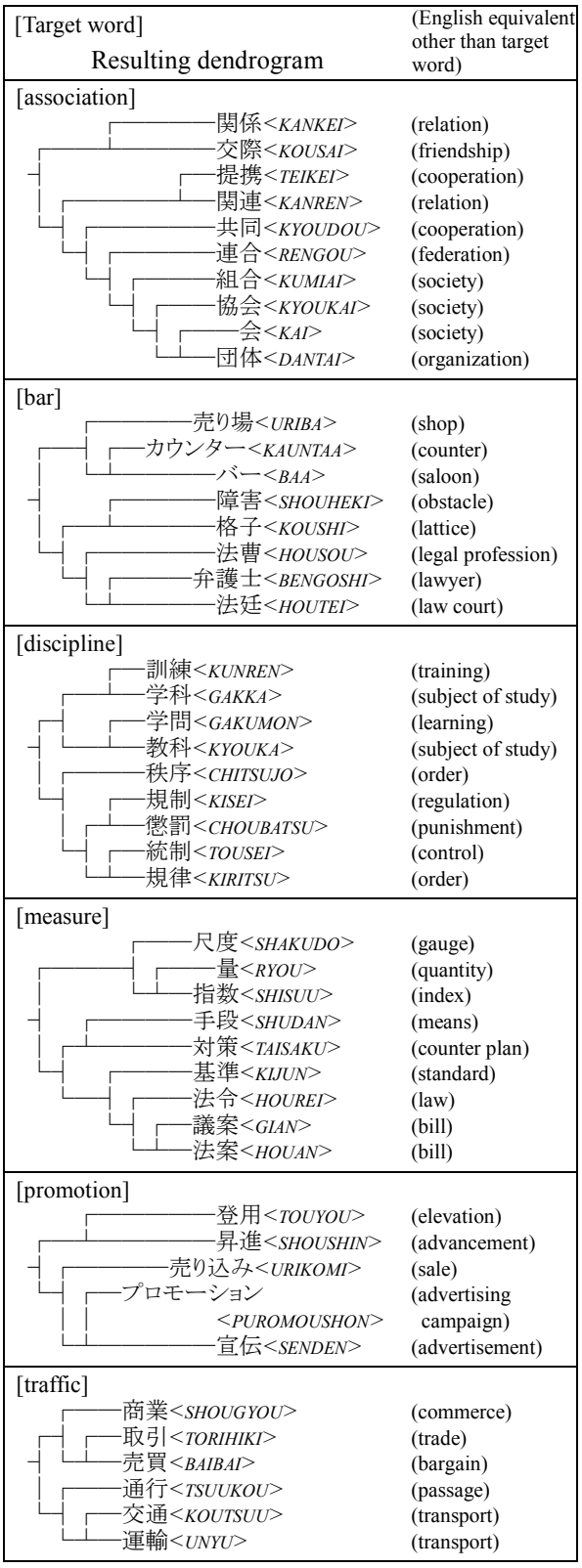
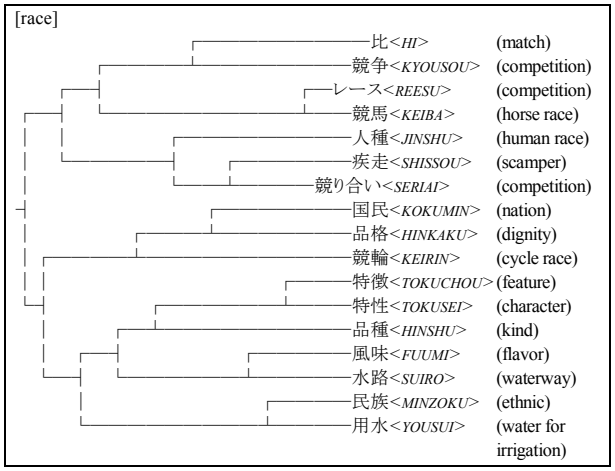
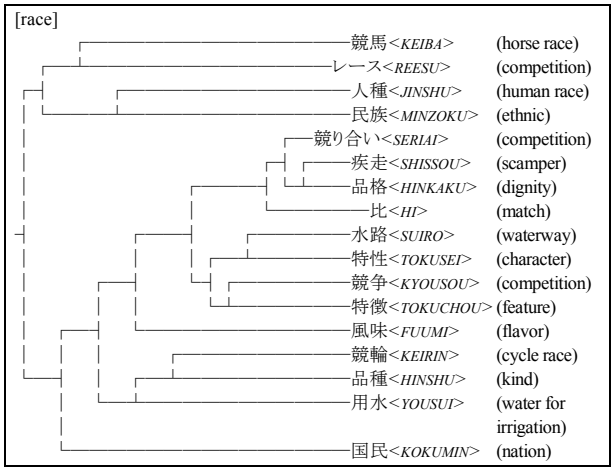


Figure 4. Examples of Clustering.

(a) Proposed method



(b) Monolingual distributional clustering



(c) Naive translingual distributional clustering

Figure 5. Comparison with Alternatives.

A Japanese *Katakana* word resulting from transliteration of an English word sometimes represents multiple senses of the English word. It is necessary to detect and split translation equivalents representing more than one sense of the target word.

Third, not only are acquired senses rather coarse-grained but also generic senses are difficult to acquire. One of the reasons for this may be that we rely on co-occurrence in the window. The fact that most distributional word clustering methods use syntactic co-occurrence suggests that it is the most effective tool for extracting pairs of related words.

6 Conclusion

We presented a translingual distributional word clustering method enabling word senses, exactly a hierarchy of clusters of translation equivalents, to be acquired from a comparable corpus and a bilingual dictionary. Its effectiveness was demonstrated through an experiment using Wall Street Journal and Nihon Keizai Shimbun corpora and the EDR bilingual dictionary. The recall of senses was 87% for senses whose ratios in the corpus were not less than 5%, and the accuracy of sense definitions was 77%.

Acknowledgments: This research was supported by the New Energy and Industrial Technology Development Organization of Japan.

References

Brown, Peter F., Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1991. Word-sense disambiguation using statistical methods. In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, pages 264-270.

Dagan, Ido and Alon Itai. 1994. Word sense disambiguation using a second language monolingual corpus. *Computational Linguistics*, 20(4): 563-596.

Fukumoto, Fumiyo and Junichi Tsujii. 1994. Automatic recognition of verbal polysemy. In *Proceedings of the 15th International Conference on Computational Linguistics*, pages 762-768.

Grefenstette, Gregory. 1994. *Explorations in Automatic Thesaurus Discovery*. Kluwer Academic Publishers, Boston.

Hindle, Donald. 1990. Noun classification from predicate-argument structures. In *Proceedings of the 28th Annual Meeting of the Association for Computational Linguistics*, pages 268-275.

Ide, Nancy and Jean Veronis. 1998. Introduction to the special issue on word sense disambiguation: The state of the art. *Computational Linguistics*, 24(1): 1-40.

Kaji, Hiroyuki and Yasutsugu Morimoto. 2002. Unsupervised word sense disambiguation using bilingual comparable corpora. In *Proceedings of the 19th International Conference on Computational Linguistics*, pages 411-417.

Pantel, Patrick and Dekang Lin. 2002. Discovering word senses from text. In *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 613-619.

Pereira, Fernando, Naftali Tishby, and Lillian Lee. 1993. Distributional clustering of English words. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, pages 183-190.

Schuetze, Hinrich. 1998. Automatic word sense discrimination. *Computational Linguistics*, 24(1): 97-124.