Selection of Effective Contextual Information for Automatic Synonym Acquisition

Masato Hagiwara, Yasuhiro Ogawa, and Katsuhiko Toyama
Graduate School of Information Science,
Nagoya University
Furo-cho, Chikusa-ku, Nagoya, JAPAN 464-8603
{hagiwara, yasuhiro, toyama}@kl.i.is.nagoya-u.ac.jp

Abstract

Various methods have been proposed for automatic synonym acquisition, as synonyms are one of the most fundamental lexical knowledge. Whereas many methods are based on contextual clues of words, little attention has been paid to what kind of categories of contextual information are useful for the purpose. This study has experimentally investigated the impact of contextual information selection, by extracting three kinds of word relationships from corpora: dependency, sentence co-occurrence, and proximity. The evaluation result shows that while dependency and proximity perform relatively well by themselves, combination of two or more kinds of contextual information gives more stable performance. We've further investigated useful selection of dependency relations and modification categories, and it is found that modification has the greatest contribution, even greater than the widely adopted subject-object combination.

1 Introduction

Lexical knowledge is one of the most important resources in natural language applications, making it almost indispensable for higher levels of syntactical and semantic processing. Among many kinds of lexical relations, synonyms are especially useful ones, having broad range of applications such as query expansion technique in information retrieval and automatic thesaurus construction.

Various methods (Hindle, 1990; Lin, 1998; Hagiwara et al., 2005) have been proposed for synonym acquisition. Most of the acquisition methods are based on distributional hypothesis (Harris, 1985), which states that semantically similar words share similar contexts, and it has been experimentally shown considerably plausible.

However, whereas many methods which adopt the hypothesis are based on contextual clues concerning words, and there has been much consideration on the language models such as Latent Semantic Indexing (Deerwester et al., 1990) and Probabilistic LSI (Hofmann, 1999) and synonym acquisition method, almost no attention has been paid to what kind of categories of contextual information, or their combinations, are useful for word featuring in terms of synonym acquisition.

For example, Hindle (1990) used co-occurrences between verbs and their subjects and objects, and proposed a similarity metric based on mutual information, but no exploration concerning the effectiveness of other kinds of word relationship is provided, although it is extendable to any kinds of contextual information. Lin (1998) also proposed an information theory-based similarity metric, using a broad-coverage parser and extracting wider range of grammatical relationship including modifications, but he didn't further investigate what kind of relationships actually had important contributions to acquisition, either. The selection of useful contextual information is considered to have a critical impact on the performance of synonym acquisition. This is an independent problem from the choice of language model or acquisition method, and should therefore be examined by itself.

The purpose of this study is to experimentally investigate the impact of contextual information selection for automatic synonym acquisition. Because nouns are the main target of

Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL, pages 353–360,
Sydney, July 2006.

synonym acquisition, here we limit the target of acquisition to nouns, and firstly extract the co-occurrences between nouns and three categories of contextual information — dependency, sentence co-occurrence, and proximity — from each of three different corpora, and the performance of individual categories and their combinations are evaluated. Since dependency and modification relations are considered to have greater contributions in contextual information and in the dependency category, respectively, these categories are then broken down into smaller categories to examine the individual significance.

Because the consideration on the language model and acquisition methods is not the scope of the current study, widely used vector space model (VSM), tf·idf weighting scheme, and cosine measure are adopted for similarity calculation. The result is evaluated using two automatic evaluation methods we proposed and implemented: discrimination rate and correlation coefficient based on the existing thesaurus WordNet.

This paper is organized as follows: in Section 2, three kinds of contextual information we use are described, and the following Section 3 explains the synonym acquisition method. In Section 4 the evaluation method we employed is detailed, which consists of the calculation methods of reference similarity, discrimination rate, and correlation coefficient. Section 5 provides the experimental conditions and results of contextual information selection, followed by dependency and modification selection. Section 6 concludes this paper.

2 Contextual Information

In this study, we focused on three kinds of contextual information: dependency between words, sentence co-occurrence, and proximity, that is, co-occurrence with other words in a window, details of which are provided the following sections.

2.1 Dependency

The first category of the contextual information we employed is the dependency between words in a sentence, which we suppose is most commonly used for synonym acquisition as the context of words. The dependency here includes predicate-argument structure such as subjects and objects of verbs, and modifications of nouns. As the extraction of accurate and comprehensive grammatical relations is in itself a difficult task, the sophisticated parser RASP Toolkit (Briscoe and Carroll, 2002) was utilized to extract this kind of word relations. RASP analyzes input sentences and provides wide variety of grammatical information such as POS tags, dependency structure, and parsed trees as output, among which we paid attention to dependency structure called grammatical relations (GRs) (Briscoe et al., 2002).

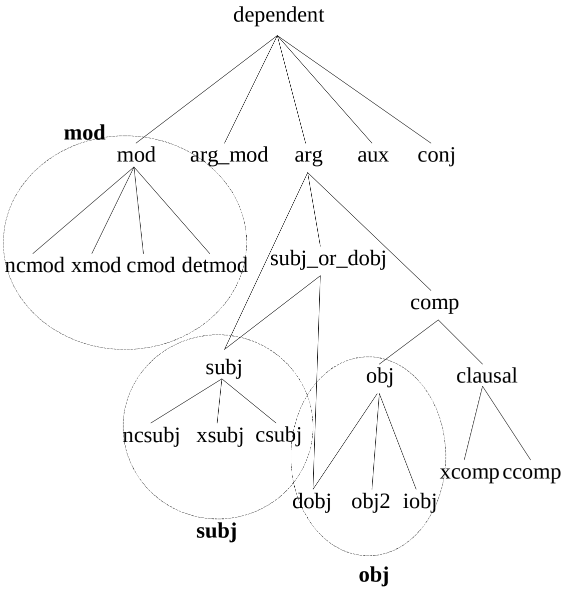


Figure 1: Hierarchy of grammatical relations and groups

GRs represent relationship among two or more words and are specified by the labels, which construct the hierarchy shown in Figure 1. In this hierarchy, the upper levels correspond to more general relations whereas the lower levels to more specific ones. Although the most general relationship in GRs is “dependent”, more specific labels are assigned whenever possible. The representation of the contextual information using GRs is as follows. Take the following sentence for example:

> Shipments have been relatively level since January, the Commerce Department noted.

RASP outputs the extracted GRs as n-ary relations as follows:

```
(ncsubj note Department obj)
(ncsubj be Shipment _)
(xcomp _ be level)
(mod _ level relatively)
(aux _ be have)
(ncmod since be January)
(mod _ Department note)
(ncmod _ Department Commerce)
```

```
(detmod _ Department the)
(ncmod _ be Department)
```

While most of GRs extracted by RASP are binary relations of head and dependent, there are some relations that contain additional slot or extra information regarding the relations, as shown "ncsubj" and "ncmod" in the above example. To obtain the final representation that we require for synonym acquisition, that is, the co-occurrence between words and their contexts, these relationships must be converted to binary relations, i.e., co-occurrence. We consider the concatenation of all the rest of the target word as context:

```
Department    ncsubj:note:*:obj
shipment      ncsubj:be:*:_
January       ncmod:since:be:*
Department    mod:_:*:note
Department    ncmod:_:*:Commerce
Commerce      ncmod:_:Department:*
Department    detmod:_:*:the
Department    ncmod:_:be:*
```

The slot for the target word is replaced by "*" in the context. Note that only the contexts for nouns are extracted because our purpose here is the automatic extraction of synonymous nouns.

2.2 Sentence Co-occurrence

As the second category of contextual information, we used the sentence co-occurrence, i.e., which sentence words appear in. Using this context is, in other words, essentially the same as featuring words with the sentences in which they occur. Treating single sentences as documents, this featuring corresponds to exploiting transposed term-document matrix in the information retrieval context, and the underlying assumption is that words that commonly appear in the similar documents or sentences are considered semantically similar.

2.3 Proximity

The third category of contextual information, proximity, utilizes tokens that appear in the vicinity of the target word in a sentence. The basic assumption here is that the more similar the distribution of proceeding and succeeding words of the target words are, the more similar meaning these two words possess, and its effectiveness has been previously shown (Macro Baroni and Sabrina Bisi, 2004). To capture the word proximity, we consider a window with a certain radius, and treat the label of the word and its position within the window as context. The contexts for the previous example sentence, when the window radius is 3, are then:

```
shipment    R1:have
shipment    R2:be
shipment    R3:relatively
January     L1:since
January     L2:level
January     L3:relatively
January     R1:,
January     R2:the
January     R3:Commerce
Commerce    L1:the
Commerce    L2:,
Commerce    L3:January
Commerce    R1:Department
...
```

Note that the proximity includes tokens such as punctuation marks as context, because we suppose they offer useful contextual information as well.

3 Synonym Acquisition Method

The purpose of the current study is to investigate the impact of the contextual information selection, not the language model itself, we employed one of the most commonly used method: vector space model (VSM) and tf·idf weighting scheme. In this framework, each word is represented as a vector in a vector space, whose dimensions correspond to contexts. The elements of the vectors given by tf·idf are the co-occurrence frequencies of words and contexts, weighted by normalized idf. That is, denoting the number of distinct words and contexts as N and M, respectively,

$$\boldsymbol{w}_i = {}^t[\mathrm{tf}(w_i, c_1) \cdot \mathrm{idf}(c_1) \ ... \ \mathrm{tf}(w_i, c_M) \cdot \mathrm{idf}(c_M)], \tag{1}$$

where $\mathrm{tf}(w_i, c_j)$ is the co-occurrence frequency of word w_i and context c_j. $\mathrm{idf}(c_j)$ is given by

$$\mathrm{idf}(c_j) = \frac{\log(N/\mathrm{df}(c_j))}{\max_k \log(N/\mathrm{df}(v_k))}, \tag{2}$$

where $\mathrm{df}(c_j)$ is the number of distinct words that co-occur with context c_j.

Although VSM and tf·idf are naive and simple compared to other language models like LSI and PLSI, they have been shown effective enough for the purpose (Hagiwara et al., 2005). The similarity between two words are then calculated as the cosine value of two corresponding vectors.

4 Evaluation

This section describes the evaluation methods we employed for automatic synonym acquisition. The evaluation is to measure how similar the obtained similarities are to the "true" similarities. We firstly prepared the reference similarities from the existing thesaurus WordNet as described in Section 4.1, and by comparing the reference and obtained similarities, two evaluation measures, discrimination rate and correlation coefficient, are calculated automatically as described in Sections 4.2 and 4.3.

4.1 Reference similarity calculation using WordNet

As the basis for automatic evaluation methods, the reference similarity, which is the answer value that similarity of a certain pair of words “should take,” is required. We obtained the reference similarity using the calculation based on thesaurus tree structure (Nagao, 1996). This calculation method requires no other resources such as corpus, thus it is simple to implement and widely used.

The similarity between word sense w_i and word sense v_j is obtained using tree structure as follows. Let the depth[1] of node w_i be d_i, the depth of node v_j be d_j, and the maximum depth of the common ancestors of both nodes be $d_{\rm dca}$. The similarity between w_i and v_j is then calculated as

$$sim(w_i, v_j) = \frac{2 \cdot d_{\rm dca}}{d_i + d_j}, \qquad (3)$$

which takes the value between 0.0 and 1.0.

Figure 2 shows the example of calculating the similarity between the word senses “hill” and “coast.” The number on the side of each word sense represents the word’s depth. From this tree structure, the similarity is obtained:

$$sim(\text{“hill”}, \text{“coast”}) = \frac{2 \cdot 3}{5 + 5} = 0.6. \qquad (4)$$

The similarity between word w with senses $w_1, ..., w_n$ and word v with senses $v_1, ..., v_m$ is defined as the maximum similarity between all the pairs of word senses:

$$sim(w, v) = \max_{i,j} sim(w_i, v_j), \qquad (5)$$

whose idea came from Lin’s method (Lin, 1998).

4.2 Discrimination Rate

The following two sections describe two evaluation measures based on the reference similarity. The first one is discrimination rate (DR). DR, originally proposed by Kojima et al. (2004), is the rate (percentage) of pairs (w_1, w_2) whose degree of association between two words w_1, w_2 is successfully discriminated by the similarity derived by the method under evaluation. Kojima et al. dealt with three-level discrimination of a pair of words, that is, highly related (synonyms or nearly synonymous), moderately related (a certain degree of association), and unrelated (irrelevant). However, we omitted the moderately related level and limited the discrimination to two-level: high or none, because of the difficulty of preparing a test set that consists of moderately related pairs.

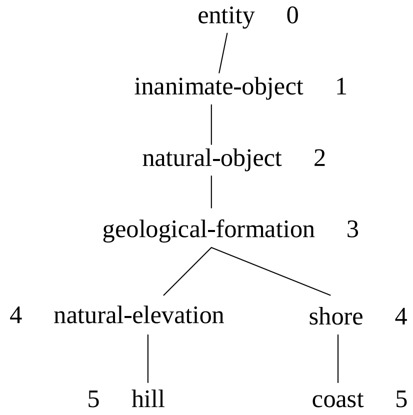


Figure 2: Example of automatic similarity calculation based on tree structure

highly related	unrelated
(answer, reply) (phone, telephone) (sign, signal) (concern, worry) ⋮	(animal, coffee) (him, technology) (track, vote) (path, youth) ⋮

Figure 3: Test-sets for discrimination rate calculation.

The calculation of DR follows these steps: first, two test sets, one of which consists of highly related word pairs and the other of unrelated ones, are prepared, as shown in Figure 3. The similarity between w_1 and w_2 is then calculated for each pair (w_1, w_2) in both test sets via the method under evaluation, and the pair is labeled highly related when similarity exceeds a given threshold t and unrelated when the similarity is lower than t. The number of pairs labeled highly related in the highly related test set and unrelated in the unrelated test set are denoted n_a and n_b, respectively.

[1] To be precise, the structure of WordNet, where some word senses have more than one parent, isn’t a tree but a DAG. The depth of a node is, therefore, defined here as the “maximum distance” from the root node.

DR is then given by:

$$\frac{1}{2}\left(\frac{n_a}{N_a}+\frac{n_b}{N_b}\right), \quad (6)$$

where N_a and N_b are the numbers of pairs in highly related and unrelated test sets, respectively. Since DR changes depending on threshold t, maximum value is adopted by varying t.

We used the reference similarity to create these two test sets. Firstly, $N_p = 100,000$ pairs of words are randomly created using the target vocabulary set for synonym acquisition. Proper nouns are omitted from the choice here because of their high ambiguity. The two testsets are then created extracting $n = 2,000$ most related (with high reference similarity) and unrelated (with low reference similarity) pairs.

4.3 Correlation coefficient

The second evaluation measure is correlation coefficient (CC) between the obtained similarity and the reference similarity. The higher CC value is, the more similar the obtained similarities are to WordNet, thus more accurate the synonym acquisition result is.

The value of CC is calculated as follows. Let the set of the sample pairs be P_s, the sequence of the reference similarities calculated for the pairs in P_s be $\boldsymbol{r} = (r_1, r_2, ..., r_n)$, the corresponding sequence of the target similarity to be evaluated be $\boldsymbol{r} = (s_1, s_2, ..., s_n)$, respectively. Correlation coefficient ρ is then defined by:

$$\rho = \frac{\frac{1}{n}\sum_{i=1}^{n}(r_i - \bar{r})(s_i - \bar{s})}{\sigma_r \sigma_s}, \quad (7)$$

where $\bar{r}, \bar{s}, \sigma_r$, and σ_s represent the average of $\boldsymbol{r}$ and $\boldsymbol{s}$ and the standard deviation of $\boldsymbol{r}$ and $\boldsymbol{s}$, respectively. The set of the sample pairs P_s is created in a similar way to the preparation of highly related test set used in DR calculation, except that we employed $N_p = 4,000, n = 2,000$ to avoid extreme nonuniformity.

5 Experiments

Now we desribe the experimental conditions and results of contextual information selection.

5.1 Condition

We used the following three corpora for the experiment: (1) Wall Street Journal (WSJ) corpus (approx. 68,000 sentences, 1.4 million tokens), (2) Brown Corpus (BROWN) (approx. 60,000 sentences, 1.3 million tokens), both of which are contained in Treebank 3 (Marcus, 1994), and (3) written sentences in WordBank (WB) (approx. 190,000 sentences, 3.5 million words) (HyperCollins, 2002). No additional annotation such as POS tags provided for Treebank was used, which means that we gave the plain texts stripped off any additional information to RASP as input.

To distinguish nouns, using POS tags annotated by RASP, any words with POS tags APP, ND, NN, NP, PN, PP were labeled as nouns. The window radius for proximity is set to 3. We also set a threshold t_f on occurrence frequency in order to filter out any words or contexts with low frequency and to reduce computational cost. More specifically, any words w such that $\sum_c \text{tf}(w, c) < t_f$ and any contexts c such that $\sum_w \text{tf}(w, c) < t_f$ were removed from the co-occurrence data. t_f was set to $t_f = 5$ for WSJ and BROWN, and $t_f = 10$ for WB in Sections 5.2 and 5.3, and $t_f = 2$ for WSJ and BROWN and $t_f = 5$ for WB in Section 5.4.

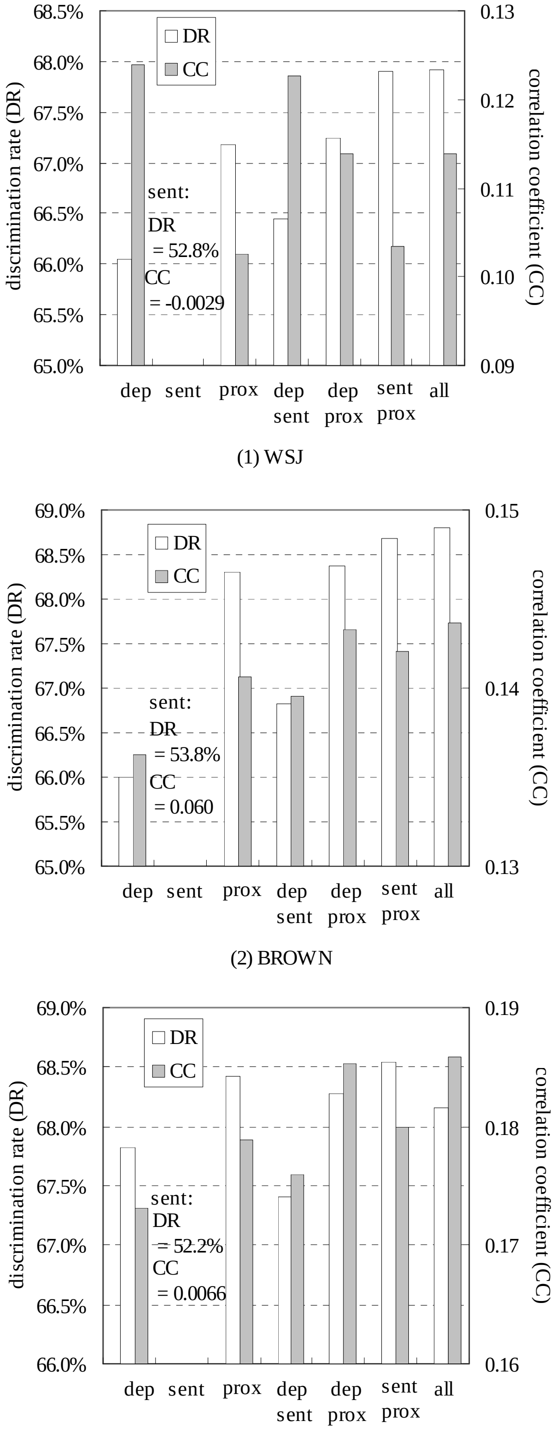
5.2 Contextual Information Selection

In this section, we experimented to discover what kind of contextual information extracted in Section 2 is useful for synonym extraction. The performances, i.e. DR and CC are evaluated for each of the three categories and their combinations.

The evaluation result for three corpora is shown in Figure 4. Notice that the range and scale of the vertical axes of the graphs vary according to corpus. The result shows that dependency and proximity perform relatively well alone, while sentence co-occurrence has almost no contributions to performance. However, when combined with other kinds of context information, every category, even sentence co-occurrence, serves to “stabilize” the overall performance, although in some cases combination itself decreases individual measures slightly. It is no surprise that the combination of all categories achieves the best performance. Therefore, in choosing combination of different kinds of context information, one should take into consideration the economical efficiency and trade-off between computational complexity and overall performance stability.

5.3 Dependency Selection

We then focused on the contribution of individual categories of dependency relation, i.e. groups of grammatical relations. The following four groups



(1) WSJ

(2) BROWN

(3) WB

Figure 4: Contextual information selection performances

Discrimination rate (DR) and correlation coefficient (CC) for (1) Wall Street Journal corpus, (2) Brown Corpus, and (3) WordBank.

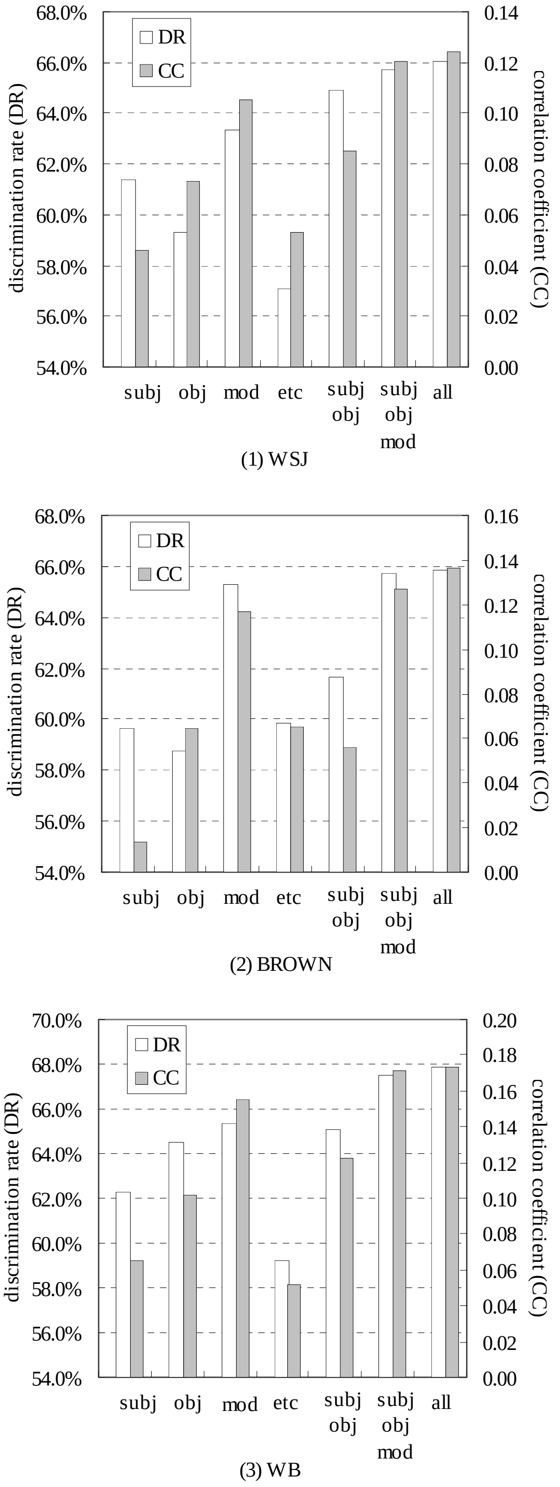
of GRs are considered for comparison convenience: (1) subj group (“subj”, “ncsubj”, “xsubj”, and “csubj”), (2) obj group (“obj”, “dobj”, “obj2”, and “iobj”), (3) mod group (“mod”, “ncmod”, “xmod”, “cmod”, and “detmod”), and (4) etc group (others), as shown in the circles in Figure 1. This is because distinction between relations in a group is sometimes unclear, and is considered to strongly depend on the parser implementation. The final target is seven kinds of combinations of the above four groups: subj, obj, mod, etc, subj+obj, subj+obj+mod, and all.

The two evaluation measures are similarly calculated for each group and combination, and shown in Figure 5. Although subjects, objects, and their combination are widely used contextual information, the performances for subj and obj categories, as well as their combination subj+obj, were relatively poor. On the contrary, the result clearly shows the importance of modification, which alone is even better than widely adopted subj+obj. The “stabilization effect” of combinations observed in the previous experiment is also confirmed here as well.

Because the size of the co-occurrence data varies from one category to another, we conducted another experiment to verify that the superiority of the modification category is simply due to the difference in the quality (content) of the group, not the quantity (size). We randomly extracted 100,000 pairs from each of mod and subj+obj categories to cancel out the quantity difference and compared the performance by calculating averaged DR and CC of ten trials. The result showed that, while the overall performances substantially decreased due to the size reduction, the relation between groups was preserved before and after the extraction throughout all of the three corpora, although the detailed result is not shown due to the space limitation. This means that what essentially contributes to the performance is not the size of the modification category but its content.

5.4 Modification Selection

As the previous experiment shows that modifications have the biggest significance of all the dependency relationship, we further investigated what kind of modifications is useful for the purpose. To do this, we broke down the mod group into these five categories according to modifying word’s category: (1) detmod, when the GR label is “det-



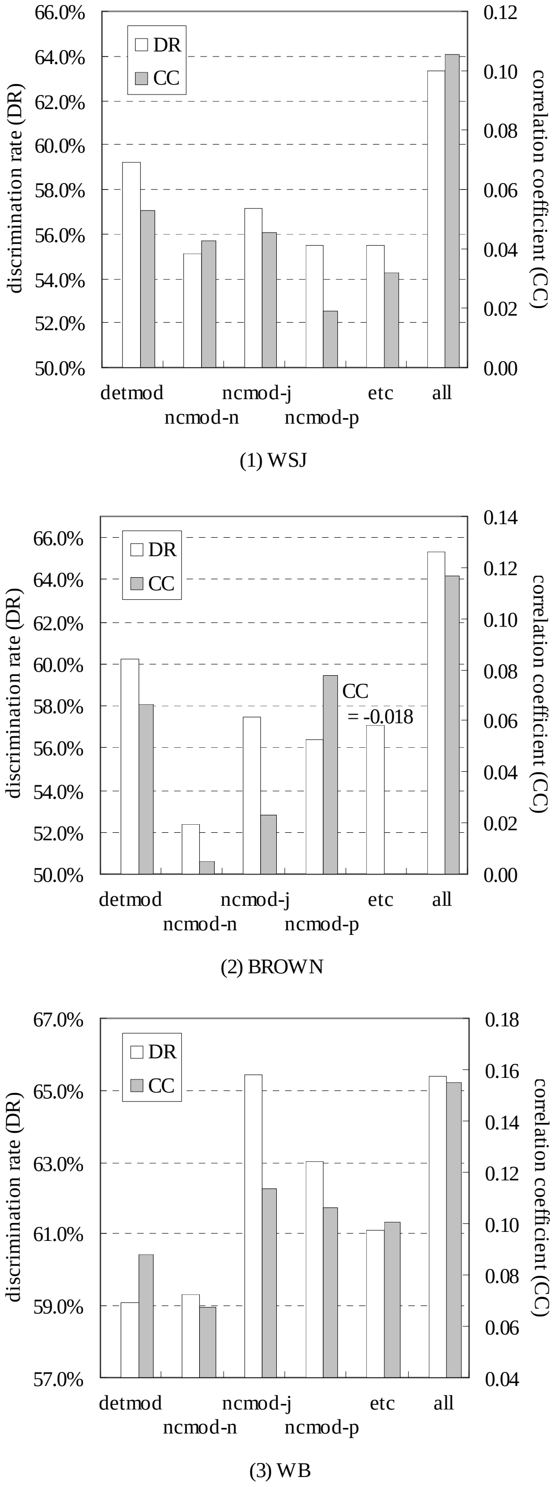
(1) WSJ

(2) BROWN

(3) WB

Figure 5: Dependency selection performances

Discrimination rate (DR) and correlation coefficient (CC) for (1) Wall Street Journal corpus, (2) Brown Corpus, and (3) WordBank.



(1) WSJ

(2) BROWN

(3) WB

Figure 6: Modification selection performances

Discrimination rate (DR) and correlation coefficient (CC) for (1) Wall Street Journal corpus, (2) Brown Corpus, and (3) WordBank.

mod”, i.e., the modifying word is a determiner, (2) ncmod-n, when the GR label is “ncmod” and the modifying word is a noun, (3) ncmod-j, when the GR label is “ncmod” and the modifying word is an adjective or number, (4) ncmod-p, when the GR label is “ncmod” and the modification is through a preposition (e.g. “state” and “affairs” in “state of affairs”), and (5) etc (others).

The performances for each modification category are evaluated and shown in Figure 6. Although some individual modification categories such as detmod and ncmod-j outperform other categories in some cases, the overall observation is that all the modification categories contribute to synonym acquisition to some extent, and the effect of individual categories are accumulative. We therefore conclude that the main contributing factor on utilizing modification relationship in synonym acquisition isn’t the type of modification, but the diversity of the relations.

6 Conclusion

In this study, we experimentally investigated the impact of contextual information selection, by extracting three kinds of contextual information — dependency, sentence co-occurrence, and proximity — from three different corpora. The acquisition result was evaluated using two evaluation measures, DR and CC using the existing thesaurus WordNet. We showed that while dependency and proximity perform relatively well by themselves, combination of two or more kinds of contextual information, even with the poorly performing sentence co-occurrence, gives more stable result. The selection should be chosen considering the trade-off between computational complexity and overall performance stability. We also showed that modification has the greatest contribution to the acquisition of all the dependency relations, even greater than the widely adopted subject-object combination. It is also shown that all the modification categories contribute to the acquisition to some extent.

Because we limited the target to nouns, the result might be specific to nouns, but the same experimental framework is applicable to any other categories of words. Although the result also shows the possibility that the bigger the corpus is, the better the performance will be, the contents and size of the corpora we used are diverse, so their relationship, including the effect of the window radius, should be examined as the future work.

References

Marco Baroni and Sabrina Bisi 2004. Using cooccurrence statistics and the web to discover synonyms in a technical language. *Proc. of the Fourth International Conference on Language Resources and Evaluation (LREC 2004).*

Ted Briscoe and John Carroll. 2002. Robust Accurate Statistical Annotation of General Text. *Proc. of the Third International Conference on Language Resources and Evaluation (LREC 2002),* 1499–1504.

Ted Briscoe, John Carroll, Jonathan Graham and Ann Copestake 2002. Relational evaluation schemes. *Proc. of the Beyond PARSEVAL Workshop at the Third International Conference on Language Resources and Evaluation,* 4–8.

Scott Deerwester, et al. 1990. Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science,* 41(6):391–407.

Christiane Fellbaum. 1998. *WordNet: an electronic lexical database.* MIT Press.

Masato Hagiwara, Yasuhiro Ogawa, Katsuhiko Toyama. 2005. PLSI Utilization for Automatic Thesaurus Construction. *Proc. of The Second International Joint Conference on Natural Language Processing (IJCNLP-05),* 334–345.

Zellig Harris. 1985. Distributional Structure. Jerrold J. Katz (ed.) *The Philosophy of Linguistics.* Oxford University Press. 26–47.

Donald Hindle. 1990. Noun classification from predicate-argument structures. *Proc. of the 28th Annual Meeting of the ACL,* 268–275.

Thomas Hofmann. 1999. Probabilistic Latent Semantic Indexing. *Proc. of the 22nd International Conference on Research and Development in Information Retrieval (SIGIR ’99),* 50–57.

Kazuhide Kojima, Hirokazu Watabe, and Tsukasa Kawaoka. 2004. Existence and Application of Common Threshold of the Degree of Association. *Proc. of the Forum on Information Technology (FIT2004)* F-003.

Collins. 2002. Collins Cobuild Mld Major New Edition CD-ROM. HarperCollins Publishers.

Dekang Lin. 1998. Automatic retrieval and clustering of similar words. *Proc. of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational linguistics (COLING-ACL ’98),* 786–774.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1994. Building a large annotated corpus of English: The Penn treebank. *Computational Linguistics,* 19(2):313–330.

Makoto Nagao (ed.). 1996. *Shizengengoshori.* The Iwanami Software Science Series 15, Iwanami Shoten Publishers.