

# Context Feature Selection for Distributional Similarity

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

Distributional similarity is a widely used concept to capture the semantic relatedness of words in various NLP tasks. However, accurate similarity calculation requires a large number of contexts, which leads to impractically high computational complexity. To alleviate the problem, we have investigated the effectiveness of automatic context selection by applying feature selection methods explored mainly for text categorization. Our experiments on synonym acquisition have shown that while keeping or sometimes increasing the performance, we can drastically reduce the unique contexts up to 10% of the original size. We have also extended the measures so that they cover context categories. The result shows a considerable correlation between the measures and the performance, enabling the automatic selection of effective context categories for distributional similarity.

## 1 Introduction

Semantic similarity of words is one of the most important lexical knowledge for NLP tasks including word sense disambiguation and synonym acquisition. To measure the semantic relatedness of words, a concept called *distributional similarity* has been widely used. Distributional similarity represents the relatedness of two words by the commonality of contexts the words share, based on the *distributional hypothesis* (Harris, 1985), which states that semantically similar words share similar contexts.

A wide range of contextual information, such as surrounding words (Lowe and McDonald, 2000; Curran and Moens, 2002a), dependency or case structure (Hindle, 1990; Ruge, 1997; Lin, 1998), and dependency path (Lin and Pantel, 2001; Pado and Lapata, 2007), has been utilized for similarity calculation, and achieved considerable success. However, a major problem which arises when adopting distributional similarity is that it easily yields a huge amount of unique contexts. This can lead to high dimensionality of context space, often up to the order of tens or hundreds of thousands, which makes the calculation computationally impractical. Because not all of the contexts are useful, it is strongly required for the efficiency to eliminate the unwanted contexts to ease the expensive cost.

To tackle this issue, Curran and Moens (2002b) suggest assigning an index vector of *canonical attributes*, i.e., a small number of representative elements extracted from the original vector, to each word. When the comparison is performed, canonical attributes of two target words are firstly consulted, and the original vectors are referred to only if the attributes have a match between them. However, it is not clear whether the condition for canonical attributes they adopted, i.e., that the attributes must be the most weighted subject, direct object, or indirect object, is optimal in terms of the performance.

There are also some existing studies which paid attention to the comparison of context categories for synonym acquisition (Curran and Moens, 2002a; Hagiwara et al., 2006). However, they have conducted only a posteriori comparison based on performance evaluation, and we are afraid that these find-

ings are somewhat limited to their own experimental settings which may not be applicable to completely new settings, e.g., one with a new set of contexts extracted from different sources. Therefore, general quantitative measures which can be used for reduction and selection of any kind of contexts and context categories are strongly required.

Shifting our attention from word similarity to other areas, a great deal of studies on feature selection has been conducted in the literature, especially for text categorization (Yang and Pedersen, 1997) and gene expression classification (Ding and Peng, 2003). Whereas these methods have been successful in reducing feature size while keeping classification performance, the problem of distributional similarity is radically different from that of classification, and whether the same methods are applicable and effective for automatic context selection in the similarity problem is yet to be investigated.

In this paper, we firstly introduce existing quantitative methods for feature selection, namely, DF, TS, MI, IG, CHI2, and show how to apply them to the distributional similarity problem to measure the context importance. We then extracted dependency relations as context from the corpus, and conducted automatic synonym acquisition experiments to evaluate the context selection performance, reducing the unimportant contexts based on the feature selection methods. Finally we extend the context importance to cover context categories (RASP2 grammatical relations), and show that the above methods are also effective in selecting categories.

This paper is organized as follows: in Section 2, five existing context selection methods are introduced, and how to apply classification-based selection methods to distributional similarity is described. In Section 3 and 4, the synonym acquisition method and evaluation measures, AP and CC, employed in the evaluation experiments are detailed. Section 5 includes two main experiments and their results: context reduction and context category selection, along with experimental settings and discussions. Section 6 concludes this paper.

## 2 Context Selection Methods

In this section, context selection methods proposed for text categorization or information retrieval are

introduced. In the following,  $n$  and  $m$  represent the number of unique words and unique contexts, respectively, and  $N(w, c)$  denotes the number of co-occurrence of word  $w$  and context  $c$ .

### 2.1 Document Frequency (DF)

Document frequency (DF), commonly used for weighting in information retrieval, is the number of documents a term co-occur with. However, in the distributional similarity settings, DF corresponds to *word frequency*, i.e., the number of unique words the context co-occurs with:

$$df(c) = |\{w | N(w, c) > 0\}|.$$

The motivation of adopting DF as a context selection criterion is the assumption that the contexts shared by many words should be informative. It is to note, however, that the contexts with too high DF are not always useful, since there are some exceptions including so-called *stopwords*.

### 2.2 Term Strength (TS)

Term strength (TS), proposed by Wilbur and Sirotkin (1992) and applied to text categorization by Yang and Wilbur (1996), measures how likely a term is to appear in “similar documents,” and it is shown to achieve a successful outcome in reducing the amount of vocabulary for text retrieval. For distributional similarity, TS is defined as:

$$s(c) = P(c \in C(w_2) | c \in C(w_1)),$$

where  $(w_1, w_2)$  is a related word pair and  $C(w)$  is a set of contexts co-occurring with the word  $w$ , i.e.,  $C(w) = \{c | N(w, c) > 0\}$ .  $s(c)$  is calculated, letting  $P_H$  be a set of related word pairs, as

$$s(c) = \frac{|\{(w_1, w_2) \in P_H | c \in C(w_1) \cap C(w_2)\}|}{|\{(w_1, w_2) \in P_H | c \in C(w_1)\}|}.$$

What makes TS different from DF is that it requires a training set  $P_H$  consisting of related word pairs. We used the test set for class  $s = 1$  as  $P_H$  described in the next section.

### 2.3 Formalization of Distributional Similarity

The following methods, MI, IG, and CHI2, are radically different from the above ones, in that they are

designed essentially for “class classification” problems. Thus we formalize distributional similarity as a classification problem as described below.

First of all, we deal with word pairs, instead of words, as the targets of classification, and define features  $f_1, \dots, f_m$  corresponding to contexts  $c_1, \dots, c_m$ , for each pair. The feature  $f_j = 1$  if the two words of the pair has the context  $c_j$  in common, and  $f_j = 0$  otherwise. Then, we define target class  $s$ , so that  $s = 1$  when the pair is semantically related, and  $s = 0$  if not. These defined, distributional similarity is formalized as a binary classification problem which assigns the word pairs to the class  $s \in \{0, 1\}$  based on the features  $c_1, \dots, c_m$ . Finally, to calculate the specific values of the following feature importance measures, we prepare two test sets of related word pairs for class  $s = 1$  and unrelated ones for class  $s = 0$ . This enables us to apply existing feature selection methods designed for classification problems to the automatic context selection.

The two test sets, related and unrelated one, are prepared using the *reference sets* described in Section 4. More specifically, we created 5,000 related word pairs by extracting from synonym pairs in the reference set, and 5,000 unrelated ones by firstly creating random pairs of LDV, whose detail is described later, and then manually making sure that no related pairs are included in these random pairs.

## 2.4 Mutual Information (MI)

Mutual information (MI), commonly used for word association and co-occurrence weighing in statistical NLP, is the measure of the degree of dependence between two events. The *pointwise* MI value of feature  $f$  and class  $s$  is calculated as:

$$I(f, s) = \log \frac{P(f, s)}{P(f)P(s)}.$$

To obtain the final context importance, we combine the MI value over both of the classes as  $I_{\max}(c_j) = \max_{s \in \{0, 1\}} I(f_j, s)$ . Note that, here we employed the maximum value of pointwise MI values since it is claimed to be the best in (Yang and Pedersen, 1997), although there can be other combination ways such as weighted average.

## 2.5 Information Gain (IG)

Information gain (IG), often employed in the machine learning field as a criterion for feature importance, is the amount of gained information of an event by knowing the outcome of the other event, and is calculated as the weighted sum of the pointwise MI values over all the event combinations:

$$G(c_j) = \sum_{f_j \in \{0, 1\}} \sum_{s \in \{0, 1\}} P(f_j, s) \log \frac{P(f_j, s)}{P(f_j)P(s)}.$$

## 2.6 $\chi^2$ Statistic (CHI2)

$\chi^2$  statistic (CHI2) estimates the lack of independence between classes and features, which is equal to the summed difference of observed and expected frequency over the contingency table cells. More specifically, letting  $F_{nm}^j$  ( $n, m \in \{0, 1\}$ ) be the number of word pairs with  $f_j = n$  and  $s = m$ , and the number of all pairs be  $N$ ,  $\chi^2$  statistic is defined as:

$$\begin{aligned} \chi^2(c_j) &= \frac{N(F_{11}F_{00} - F_{01}F_{10})}{(F_{11} + F_{01})(F_{10} + F_{00})(F_{11} + F_{10})(F_{01} + F_{00})}. \end{aligned}$$

## 3 Synonym Acquisition Method

This section describes the synonym acquisition method, a major and important application of distributional similarity, which we employed for the evaluation of automatic context selection. Here we mention how to extract the original contexts from corpora in detail, as well as the calculation of weight and similarity between words.

### 3.1 Context Extraction

We adopted dependency structure as the context of words since it is the most widely used and well-performing contextual information in the past studies (Ruge, 1997; Lin, 1998). As the extraction of accurate and comprehensive dependency structure is in itself a difficult task, the sophisticated parser RASP Toolkit 2 (Briscoe et al., 2006) was utilized to extract this kind of word relations. Take the following sentence for example:

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

RASP outputs the extracted dependency structure as n-ary relations as follows, which are called *grammatical relations*. Annotations regarding suffix, part of speech tags, offsets for individual words are omitted for simplicity.

```
(ncsubj be Shipment _)
(aux be have)
(xcomp _ be level)
(ncmod _ be relatively)
(ccomp _ level note)
(ncmod _ note since)
(ncsubj note Department _)
(det Department the)
(ncmod _ Department Commerce)
(dobj since January)
```

While the RASP outputs are n-ary relations in general, what we need here is co-occurrences of words and contexts, so we extract the set of co-occurrences of stemmed words and contexts by taking out the target word from the relation and replacing the slot by an asterisk “\*”:

```
(words)      - (contexts)
Shipment     - ncsbj:be:*_
have        - aux:be:*
be          - ncsbj:*:Shipment:_
be          - aux:*:have
be          - xcomp:_:*:level
be          - ncmod:_:*:relatively
relatively  - ncmod:_:be:*
level       - xcomp:_:be:*
level       - ccomp:_:*:note
...
```

Summing all these up produces the raw co-occurrence count  $N(w, c)$  of word  $w$  and context  $c$ .

### 3.2 Similarity Calculation

Although it is possible to use the raw count acquired above for the similarity calculation, directly using the raw count may cause performance degradation, thus we need an appropriate weighting measure. In response to the preliminary experiment results, we employed pointwise mutual information as weight:

$$\text{wgt}(w, c) = \log \frac{P(w, c)}{P(w)P(c)}$$

Here we made a small modification to bind the weight to non-negative such that  $\text{wgt}(w, c) \geq 0$ , because negative weight values sometimes worsen the performance (Curran and Moens, 2002b). The weighting by PMI is applied *after* the pre-processing including frequency cutoff and context selection.

As for the similarity measure, we used Jaccard coefficient, which is widely adopted to capture overlap proportion of two sets:

$$\frac{\sum_{c \in C(w_1) \cap C(w_2)} \min(\text{wgt}(w_1, c), \text{wgt}(w_2, c))}{\sum_{c \in C(w_1) \cup C(w_2)} \max(\text{wgt}(w_1, c), \text{wgt}(w_2, c))}$$

## 4 Evaluation Measures

This section describes the two evaluation methods we employed — average precision (AP) and correlation coefficient (CC).

### 4.1 Average Precision (AP)

The first evaluation measure, average precision (AP), is a common evaluation scheme for information retrieval, which evaluates how accurately the methods are able to extract synonyms. We first prepare a set of *query words*, for which synonyms are obtained to evaluate the precision. We adopted the Longman Defining Vocabulary (LDV) <sup>1</sup> as the candidate set of query words. For each word in LDV, three existing thesauri are consulted: Roget’s Thesaurus (Roget, 1995), Collins COBUILD Thesaurus (Collins, 2002), and WordNet (Fellbaum, 1998). The union of synonyms obtained when the LDV word is looked up as a noun is used as the *reference set*, except for words marked as “idiom,” “informal,” “slang” and phrases comprised of two or more words. The LDV words for which no noun synonyms are found in any of the reference thesauri are omitted. From the remaining 771 LDV words, 100 query words are randomly extracted, and for each of them the eleven precision values at 0%, 10%, ..., and 100% recall levels are averaged to calculate the final AP value.

### 4.2 Correlation Coefficient (CC)

The second evaluation measure is correlation coefficient (CC) between the target similarity and the *reference similarity*, i.e., the answer value of similarity for word pairs. The reference similarity is calculated based on the closeness of two words in the tree structure of WordNet. More specifically, the similarity between word  $w$  with senses  $w_1, \dots, w_{m_1}$  and word  $v$  with senses  $v_1, \dots, v_{m_2}$  is obtained as follows. Let the depth of node  $w_i$  and  $v_j$  be  $d_i$  and  $d_j$ ,

<sup>1</sup>[http://www.cs.utexas.edu/users/kbarker/working\\_notes/ldoce-vocab.html](http://www.cs.utexas.edu/users/kbarker/working_notes/ldoce-vocab.html)

and the depth of the deepest common ancestors of both nodes be  $d_{dca}$ . The similarity is then

$$sim(w, v) = \max_{i,j} sim(w_i, v_j) = \max_{i,j} \frac{2 \cdot d_{dca}}{d_i + d_j},$$

which takes the value between 0.0 and 1.0. Then, the value of CC is calculated as the correlation coefficient of reference similarities  $\mathbf{r} = (r_1, r_2, \dots, r_n)$  and target similarities  $\mathbf{s} = (s_1, s_2, \dots, s_n)$  over the word pairs in sample set  $P_s$ , which is created by choosing the most similar 2,000 word pairs from 4,000 randomly created pairs from LDV. To avoid test-set dependency, all the CC values presented in this paper are the average values of three trials using different test sets.

## 5 Experiments

Now we describe the experimental settings and the evaluation results of context selection methods.

### 5.1 Experimental Settings

As for the corpus, New York Times section of English Gigaword <sup>2</sup>, consisting of around 914 million words and 1.3 million documents was analyzed to obtain word-context co-occurrences. Frequency cut-off was applied as a pre-processing in order to filter out any words and contexts with low frequency and to reduce computational cost. More specifically, any words  $w$  such that  $\sum_c tf(w, c) < \theta_f$  and any contexts  $c$  such that  $\sum_w tf(w, c) < \theta_f$ , with  $\theta_f = 40$ , were removed from the co-occurrence data.

Since we set our purpose here to the automatic acquisition of synonymous nouns, only the nouns except for proper nouns were selected. To distinguish nouns, using POS tags annotated by RASP2, any words with POS tags APP, ND, NN, NP, PN, PP were labeled as nouns. This left a total of 40,461 unique words and 139,618 unique context, which corresponds to the number of vectors and the dimensionality of semantic space, respectively.

### 5.2 Context Reduction

In the first experiment, we show the effectiveness of the five contextual selection methods introduced in Section 2 for context reduction problem. The five

<sup>2</sup><http://www ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2003T05>

measures were calculated for each context, and contexts were sorted by their importance. The change of performance, AP and CC, was calculated on eliminating the low-ranked contexts and varying the proportion of remaining ones, until only 0.2% (279 in number) of the unique contexts are left.

The result is displayed in Figure 1. The overall observation is that the performance not only kept the original level but also slightly improved even during the “aggressive” reduction when more than 80% of the original contexts were eliminated and less than 20,000 contexts were left. It was not until 90% (approx. 10,000 remaining) elimination that the AP values began to fall. The tendency of performance change was almost the same for AP and CC, but we observe a slight difference regarding which of the five measures were effective. More specifically, TS, IG and CHI2 worked well for AP, and DF, TS, while CHI2 did for CC. On the whole, TS and CHI2 were performing the best, whereas the performance of MI quickly worsened. Although the task is different, this experiment showed a very consistent result compared with the one of Yang and Pedersen’s (1997). This means that feature selection methods are also effective for context selection in distributional similarity, and our formalization of the problem described in Section 2 turned out to be appropriate for the purpose.

### 5.3 Context Category Selection

We are then naturally interested in what kinds of contexts are included in these top-ranked effective ones and how much they affect the overall performance. To investigate this, we firstly built a set of *elite contexts*, by gathering each top 10% (13,961 in number) contexts chosen by DF, TS, IG, and CHI2, and obtaining the intersection of these four top-ranked contexts. It was found that these four had a great deal of overlap among them, the number of which turned out to be 6,440.

Secondly, to measure the degree of effect a context category has, we defined *category importance* as the sum of all IG values of the contexts which belong to the category. The reason is that, (a) IG was one of the best-performing criteria as the previous experiment showed, and (b) IG value for a set of contexts can be calculated as the sum of IG values of individual elements, assuming that all the contexts

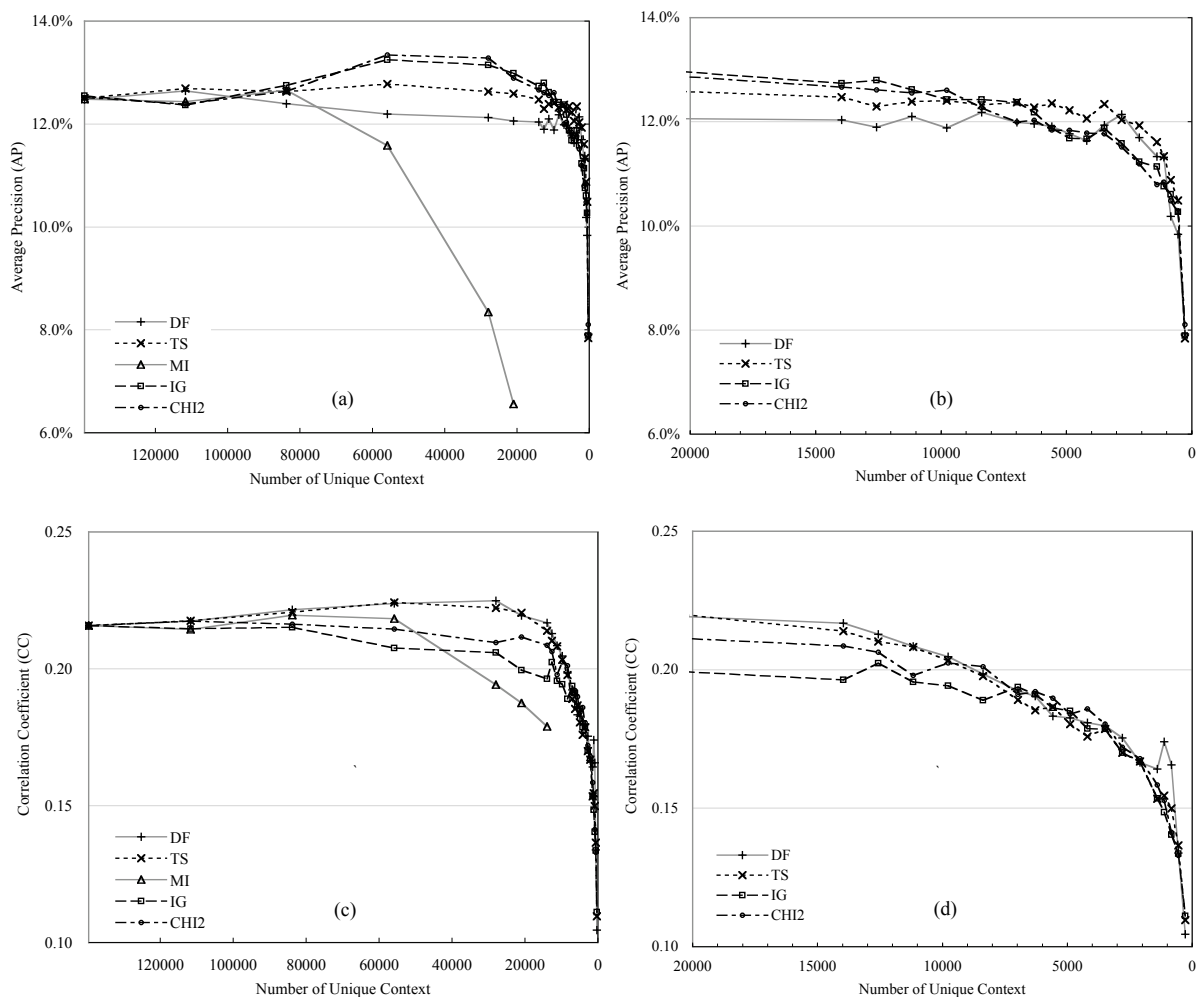


Figure 1: Performance of synonym acquisition on automatic context reduction

(a) The overall view and (b) the close-up of 0 to 20,000 unique contexts for AP, and (c) the overall view and (d) the close-up for CC

are mutually independent, which is a naive but practical assumption because of the high independence of acquired contexts from corpora.

For the categories: *ncsubj*, *dobj*, *obj*, *obj2*, *ncmod*, *xmod*, *cmod*, *ccomp*, *det*, *ta*, based on the RASP2 grammatical relations which occur frequently (more than 1.0%) in the corpus, their category importance within the elite context set was computed and showed in Figure 2. The graph also shows the performance of individual context categories, calculated when each category was separately extracted from the entire corpus. The result indicates that there is a considerable correlation

( $r = 0.760$ ) between category importance and performance, which means it is possible to predict the final performance of any context categories by calculating their category importance values in the limited size of selected context set.

As for the qualitative difference of category types, the result also shows the effectiveness of modification (*ncmod*) category, which is consistent with the result (Hagiwara et al., 2006) that *mod* is more contributing than *subj* and *obj*, which have been extensively used in the past. However, it can be seen that the reason why the *ncmod* performs well may be only because it is the largest category in size (2,515

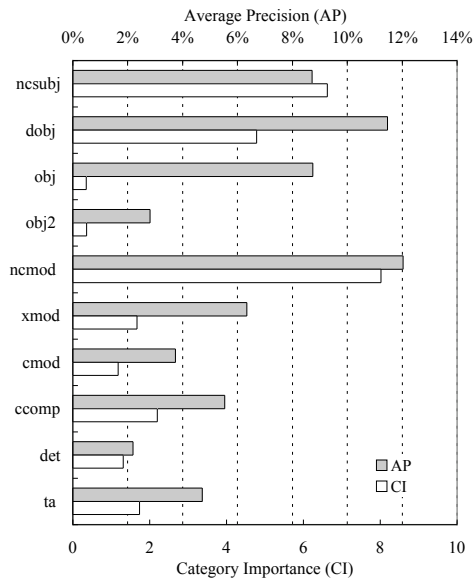


Figure 2: Performance of synonym acquisition vs context category importance

in the elite contexts). The investigation of the relations between context size and performance should be conducted in the future.

## 6 Conclusion

In this study, we firstly introduced feature selection methods, previously proposed for text categorization, and showed how to apply them for automatic context selection for distributional similarity by formalizing the similarity problem as classification. We then extracted dependency-based context from the corpus, and conducted evaluation experiments on automatic synonym acquisition.

The experimental results showed that while keeping or even improving the original performance, it is possible to eliminate a large proportion of contexts (almost up to 90%). We also extended the context importance to cover context categories based on RASP2 grammatical relations, and showed a considerable correlation between the importance and the actual performance, suggesting the possibility of automatic context category selection.

As the future works, we should further discuss other kinds of formalization of distributional similarity and their impact, because we introduced and

only briefly described a quite simple formalization model in Section 2.3. More detailed investigations on the contributions of sub-categories of contexts, and other contexts than dependency structure, such as surrounding words and dependency path, is also the future work.

## References

- Ted Briscoe, John Carroll and Rebecca Watson. 2006. The Second Release of the RASP System. *Proc. of the COLING/ACL 2006 Interactive Presentation Sessions*, 77–80.
- Collins. 2002. Collins Cobuild Major New Edition CD-ROM. HarperCollins Publishers.
- James R. Curran and Marc Moens. 2002. Scaling Context Space. *Proc. of ACL 2002*, 231–238.
- James R. Curran and Marc Moens. 2002. Improvements in automatic thesaurus extraction. In Workshop on Unsupervised Lexical Acquisition. *Proc. of ACL SIGLEX*, 231–238.
- Chris Ding and Hanchuan Peng. 2003. Minimum Redundancy Feature Selection from Microarray Gene Expression Data. *Proc. of the IEEE Computer Society Conference on Bioinformatics*, 523–528.
- Editors of the American Heritage Dictionary. 1995. Roget’s II: The New Thesaurus, 3rd ed. *Houghton Mifflin*.
- Christiane Fellbaum. 1998. *WordNet: an electronic lexical database*, MIT Press.
- Masato Hagiwara, Yasuhiro Ogawa, Katsuhiko Toyama. 2006. Selection of Effective Contextual Information for Automatic Synonym Acquisition. *Proc. of COLING/ACL 2006*, 353–360.
- 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.
- Will Lowe and Scott McDonald. 2000. The direct route: Mediated priming in semantic space. *Proc. of the 22nd Annual Conference of the Cognitive Science Society*, 675–680.
- Dekang Lin. 1998. Automatic retrieval and clustering of similar words. *Proc. of COLING/ACL 1998*, 786–774.

- Dekang Lin and Patrick Pantel. 2001. Discovery of inference rules for question answering. *Natural Language Engineering*, Volume 7, Issue 4, 343–360.
- Seastian Pado and Mirella Lapata. 2007. Dependency-Based Construction of Semantic Space Models. *Computational Linguistics*, Volume 33, Issue 2, 161–199.
- Gerda Ruge. 1997. Automatic detection of thesaurus relations for information retrieval applications. *Foundations of Computer Science: Potential - Theory - Cognition*, LNCS, Volume 1337, 499–506, Springer Verlag, Berlin, Germany.
- Yiming Yang and John Wilbur. 1996. Using corpus statistics to remove redundant words in text categorization. *Journal of the American Society for Information Science*, Volume 47, Issue 5, 357–369.
- Yiming Yang and Jan O. Pedersen. 1997. A Comparative Study on Feature Selection in Text Categorization. *Proc. of ICML 97*, 412–420.
- John Wilbur and Karl Sirotkin. 1992. The automatic identification of stop words. *Journal of Information Science*, 45–55.