A Method for Similarity-based Lexical Disambiguation

Qujiang Peng,[†] Haodong Wu^{††} and Teiji Furugori[†]

This paper describes a method for word sense disambiguation using a similarity metric. In this method, we first obtain context-similarity vectors for the senses of a polysemous word using a corpus and also define the context representation for the same polysemous word appearing in text. We then calculate distributional matrix between each context-similarity vector and the context representation for the word to be disambiguated. Finally, comparing the values of distributional matrices, we select the sense with the highest value as the meaning of the polysemous word. An experiment with 682 instances for 10 polysemous words shows that we are able to disambiguate at a rate of almost 92%.

Keywords: *Ambiguity, Word Sense, Disambiguation, Similarity, Context, Distributional Matrix, Density*

1 Introduction

The meaning of a word is not unique. This makes natural languages nothing but a convenient and flexible tool to use for communication. However, it is a big problem in natural language processing. To understand a passage like,

Water the plant in the garden.

we need to know what *plant* means: a living thing or factory? Unless this problem is resolved, we will have no better systems, not to mention a complete one, for machine translation, natural language understanding, information retrieval, etc.

In what follows, we present a method and its experimental result for resolving lexical ambiguities based on a word similarity measure.

2 Word Sense Disambiguation

The history tells us that early work in lexical or word sense disambiguation (WSD) took a pure AI approach (e.g., Hirst 1987; Small and Rieger 1982). Many researchers manually set up a knowledge base for each word and described its senses in various linguistic usages. Then, some started to use machine-readable dictionaries (MRDs) to select the proper sense

[†] The University of Electro-Communications, Department of Computer Science

^{††} Dokkyo University, Faculty of Foreign Languages

of a polysemous word (e.g., Guthrie, Guthrie, and Aidinejad 1991; Lesk 1986; Walker 1987).

The MRDs are widely available, but they have deficiencies for WSD. For instance, the definitions of a word in them are often too short and too uneven to cover necessary collocations. We encounter another difficulty in retrieving necessary information from the on-line dictionaries since their contents are usually expressed for the use of human beings rather than for the computers.

The research on WSD relies on corpora in the most of work being done nowadays. The corpus-based approach has an advantage in its general applicability. It is also possible to get around costly and perhaps error-prone parsing or semantic analysis otherwise needed. Along this line, Yarowsky (Yarowsky 1992) presented a statistical method of resolving lexical ambiguities with the use of Roget's Thesaurus and a large corpus. Dagan, Marcus, and Markovitch (Dagan, Marcus, and Markovitch 1995) proposed a method based on the adjacent word of a polysemous word to disambiguate the word senses. Leacock, Chodorow, and Miller (Leacock, Chodorow, and Miller 1998) determined the sense of a word in text using a corpus and WordNet (WordNet 2001) relations. Schütze (Schütze 1998) proposed a method for WSD based on clustering on contextual similarities. Karov and Edelman (Karov and Edelman 1998) devised a disambiguation method using a corpus and MRD, and attempted to determine the word senses by the different contextual usages of the polysemous word being analyzed.

3 Measurement of Similarity

Word similarity between two words w_1 and w_2 depends on the degree of their co-occurrences in text. The contextually similar words are the words co-occurring frequently in a distance in text. They need not be synonyms, or belong to the same syntactic or semantic category. For instance, the words *doctor* and *health* are not synonyms in thesauruses, but they are likely to be contextually similar as they tend to co-occur in a text.

3.1 Mutual Information

The contextually similar word may be measured by using mutual information. The mutual information, I , estimates the strength of association between two words w_1 and w_2 (Church and Hanks 1990):

$$I(w_1, w_2) = \log_2 \left(\frac{N * f(w_1, w_2)}{f(w_1)f(w_2)} \right) \quad (1)$$

Here, N in (1) is the size of the corpus used in the estimation, $f(w_1, w_2)$ is the frequency of co-occurrences of w_1 and w_2 , and $f(w_1)$ or $f(w_2)$ is the frequency of each word. If there is

a strong association between w_1 and w_2 , then $I(w_1, w_2) \gg 0$. If there is a weak association between w_1 and w_2 , then $I(w_1, w_2) \approx 0$. If $I(w_1, w_2) \ll 0$, then w_1 and w_2 are said to be in complementary distribution.

If w_1 and w_2 are contextually similar, then $I(w, w_1)$ and $I(w, w_2)$ have similar mutual information for some word w in lexicon. We will use the third word w in the next section to measure the similarity of two words w_1 and w_2 .

3.2 Similarity Metric of Two Words

Mathematically, the error can be used to express the dissimilarity of two words w_1 and w_2 . For example, $|I(w, w_1) - I(w, w_2)|$ shows the absolute error in that it indicates how much the two numbers $I(w, w_1)$ and $I(w, w_2)$ differ. The absolute error does not express the precision of the difference, however. So, we define the proportional error:

$$\frac{|I(w, w_1) - I(w, w_2)|}{I(w, w_1) + I(w, w_2)}$$

We know that the bigger the similar values of $I(w, w_1)$ and $I(w, w_2)$, the stronger the similarity of w_1 and w_2 . In the proportional error, however, $(I(w, w_1), I(w, w_2)) = (2.7, 3.2)$ and $(I(w, w_1), I(w, w_2)) = (0.27, 0.32)$, for instance, have the same effect on measuring the similarity, though in reality the numbers in the second pair are far smaller and thus far less important (sometimes harmful) in measuring the similarity.

3.2.1 Strong Bridge Set

It is desirable to measure the word similarity collecting the set of word w so that its member has bigger mutual information with w_1 and w_2 . We define such a word set as:

$$S = \left\{ w \mid I(w, w_1) > t_0, I(w, w_2) > t_0, t_1 < \frac{I(w, w_1)}{I(w, w_2)} < t_2, w \in \text{Lexicon} \right\}$$

and call it the strong bridge set. Here, t_0 , t_1 , and t_2 are threshold values, t_0 makes eliminate small mutual information values, and t_1 and t_2 make it possible to get the word w for which w_1 and w_2 have similar mutual information. We set $t_2 = t_1^{-1}$. The similarity of w_1 and w_2 gets stronger as $\|S\|$ becomes bigger.

3.2.2 Definition of Similarity

We will define a similarity metric of two words w_1 and w_2 using the proportional error and the strong bridge set S .

$$\text{sim}(w_1, w_2) = \log_{10} \sum_{w \in S} e^{-\frac{|I(w, w_1) - I(w, w_2)|}{I(w, w_1) + I(w, w_2)}} \quad (2)$$

The function $f(x) = e^{-x}$ in (2) can convert dissimilarity to similarity. The function $f(x) = \log_{10}(x)$ is to make the function value small.

Evidently, equation (2) is symmetric, i.e., $\text{sim}(w_1, w_2) = \text{sim}(w_2, w_1)$. This agrees with an intuition that the similarity of two words w_1 and w_2 is equal to that of w_2 and w_1 .

In equation (2), the word w_1 is most similar to itself. It is proven by showing $\text{sim}(w_1, w_2) \leq \text{sim}(w_1, w_1)$. Put it simply,

$$\begin{aligned} S(w_1, w_2) &= \left\{ w \mid I(w, w_1) > t_0, I(w, w_2) > t_0, t_1 < \frac{I(w, w_1)}{I(w, w_2)} < t_2, w \in \text{Lexicon} \right\} \\ &\subset \{w \mid I(w, w_1) > t_0, w \in \text{Lexicon}\} = S(w_1, w_1) \end{aligned}$$

and then we get:

$$\begin{aligned} \text{sim}(w_1, w_2) &= \log_{10} \sum_{w \in S(w_1, w_2)} e^{-\frac{|I(w, w_1) - I(w, w_2)|}{I(w, w_1) + I(w, w_2)}} \leq \log_{10} \sum_{w \in S(w_1, w_2)} e^{-0} \\ &\leq \log_{10} \sum_{w \in S(w_1, w_1)} e^{-0} = \text{sim}(w_1, w_1) \end{aligned}$$

4 A Method of Word Sense Disambiguation

We now devise a method of WSD based on equation (2). It takes the steps in Figure 1.

Figure 1 The Process of WSD

Word sense disambiguation starts with obtaining context-similarity vector, $V_{s_1}, V_{s_2}, \dots, V_{s_r}$, for each sense of a polysemous word, w , using the word similarity metric in equation (2). At the same time, we get the context representation, V_{CR} , for the same word, w , appearing in text. We then calculate the similarity, $sense(V_{s_m}, V_{CR})$, between each vector and the context representation. Finally, we compare the values of similarities, $sense(V_{s_1}, V_{CR}), sense(V_{s_2}, V_{CR}), \dots, sense(V_{s_r}, V_{CR})$, and select the sense, s_m , with the highest value as the meaning of the polysemous word in question.

4.1 Context Representation and Context-similarity Vector

The context representation (CR) we use here is the sequence of content words (nouns, verbs, adjectives) that appear in a distance l with the polysemous word in question. It is defined as the vector:

$$V_{CR} = (w_1, w_2, \dots, w_i, w_{i+1}, \dots, w_n) \quad (3)$$

Let s_m be the m th sense of a polysemous word w . Using equation (2), we can get a vector of each sense of a polysemous word.

$$V_{s_m} = (w_1^m, w_2^m, \dots, w_k^m) \quad (4)$$

Here $w_1^m, w_2^m, \dots, w_k^m$ are the set of the k most contextually similar words of the m th sense of a polysemous word w . s_m and $w_i^m (i = 1, \dots, k)$ have big and similar mutual information with some other word. We set k to be 6 in this paper.

When the m th sense of a polysemous word w in the corpus faces the data sparseness problem, we seek its synonym set in the WordNet, choose from the set a word that does not have the problem, and use it to calculate V_{s_m} . When no such word is found, we use its coordinate word (WordNet 2001) that does not have the problem to calculate V_{s_m} .

4.2 Distributional Matrix and the Meaning of Polysemous Word

Using (3) and (4), we will get a distributional matrix of mutual information.

$$M(V_{s_m}, V_{CR}) = (I(w_i^m, w_j))_{k \times n} \quad (i = 1, \dots, k; j = 1, \dots, n) \quad (5)$$

The matrix $M(V_{s_m}, V_{CR})$ expresses a distribution of the tight degree of association between the m th sense of a polysemous word w and the context V_{CR} . And we can get the density of the matrix $M(V_{s_m}, V_{CR})$.

$$\rho(V_{s_m}, V_{CR}) = \frac{\sum_{i=1}^k \sum_{j=1}^n I(w_i^m, w_j)}{k \times n} \quad (6)$$

If $I(w_i^m, w_j)$ is only a pure number, then $\rho(V_{s_m}, V_{CR})$ is only a mean value of some numbers. However, $I(w_i^m, w_j)$ has linguistic significance, i.e., it also expresses the tight degree of association between two words. Consequently, $\rho(V_{s_m}, V_{CR})$ is not only a mean value but also expresses the tight degree of association between the m th sense of a polysemous word w and the context V_{CR} .

In the context of natural language, there are noisy words or irrelevant information (Dagan and Itai 1994) in V_{CR} to express the topic involved. So we set the threshold t to modify (6) so that we select the most informative context words (Gale, Church and Yarowsky 1992) for word sense disambiguation. We throw away the mutual information of some words that have little contribution for disambiguation.

Consequently, we make the distance between densities enlarged and the difference of the senses of a polysemous word more evident. The process will raise our confidence in identifying the word sense, especially, in the case $\rho(V_{s_1}, V_{CR}) \approx \rho(V_{s_2}, V_{CR})$.

In the matrix $M(V_{s_m}, V_{CR})$, we throw away the values of mutual information in d most small values if the value is less than t . For the rest, we calculate the density of mutual information. We mark the density with $sense(V_{s_m}, V_{CR})$. The parameters d and t are adjustable.

Yarowsky (Yarowsky 1993, 1994a, 1994b) suggests that semantic or topic-based ambiguities require a large window of 20-50 words. In our experiment, we use windows of $l=10, 35$, and 50 words before and after the polysemous word to be disambiguated. When $l = 10$, we do not use t and d . For $l = 35$ and $l = 50$, we set t to be

$$t = \frac{\sum_{i=1}^k \sum_{j=1}^n I(w_i^m, w_j)}{4k \times n}$$

and d to be $9k$ and $12k$, respectively.

Suppose the word appearing in a text is w , and its lexical meanings are $s_1, s_2, \dots, s_r$. We calculate the density $sense(V_{s_1}, V_{CR}), sense(V_{s_2}, V_{CR}), \dots, sense(V_{s_r}, V_{CR})$. The meaning of the word in consideration is s_m that satisfies the maximal value.

5 Experiment and Result

To test our method, we use EDR English Corpus (EDR 1998) as training data and 10 polysemous words for which 682 instances were selected randomly from (Text1 1998; Text2 1998) and other materials. The EDR corpus contains 160,000 sentences with annotated morphological, syntactic and semantic information. And we set $t_0 = 0.5$, $t_1 = 0.75$, and $t_2 = 1.33$.

We predetermined the meaning of each instance of the polysemous word by two human

subjects. The meaning of w is determined by the following procedure:

- (1) Obtain the vectors $V_{s_1}, V_{s_2}, \dots, V_{s_r}$ for the lexical meanings of w .
- (2) Get V_{CR} in consideration using a window of l words each before and after w in the text.
- (3) Calculate the density $sense(V_{s_1}, V_{CR}), sense(V_{s_2}, V_{CR}), \dots, sense(V_{s_r}, V_{CR})$.
- (4) Select s_m that got the maximal density value to be the meaning of w .

5.1 Examples

Consider the polysemous word in consideration to be *cabinet* in the following text and see how its meaning is determined:

... Chicago was the industrial inferno of the nineteenth century A.D. A curious anecdote has come down to us of John Burns, a great English labor leader and one time member of the British *cabinet*. In Chicago, while on a visit to the United States, he was asked by a newspaper reporter for his opinion of that city. 'Chicago,' he answered, 'is a pocket edition of hell.' Some time later ... (*a text on Internet*)

Cabinet is given two nominal meanings: administrative organ (s_1) and a shelf (s_2). Using equation (2), we get for them the context-similarity vectors:

$$V_{s_1} = (\text{minister, administration, council, approve, prime, regular})$$

$$V_{s_2} = (\text{display, exhibit, store, maker, tea, shelf})$$

The context representation V_{CR} ($l = 35$) from the text is:

$$V_{CR} = (\dots, \text{time, member, British, Chicago, visit}, \dots)$$

The calculation of the distributional matrix produces the density 0.411462 for $sense(V_{s_1}, V_{CR})$ and 0.180325 for $sense(V_{s_2}, V_{CR})$. So, we decide the meaning of the *cabinet* to be s_1 (administrative organ).

Take another example for the word *sentence* ($l = 35$) in:

... and there he died like a brave man. He refused to have his eyes bandaged, saying that he was not at all afraid of death; and he admitted the justice of his *sentence*, and was much regretted by the people. Although Mary had shrunk at the most important time from disproving her guilt, she was very careful never to do anything that would admit ... (*Also a text on Internet*)

Again, *sentence* has two meanings: punishment (s_1) and group of words (s_2). The context-similarity vectors and the context representation are:

$$V_{s_1} = (\text{prison, judge, murder, trial, crime, ruling})$$

$$V_{s_2} = (\text{word, code, language, meaning, symbol, accent})$$

$$V_{CR} = (\dots, \text{death, admit, justice, regret, people}, \dots)$$

Now, the density for $sense(V_{s_1}, V_{CR})$ becomes 0.497328 and $sense(V_{s_2}, V_{CR})$ becomes 0.179224. Thus, we get the meaning of the *sentence* as s_1 (punishment).

5.2 Experimental Results

Table 1 shows the 10 polysemous words, their senses, the number of their instances, and the instances identified correctly in our experiment using $l = 35$ (Experiment 2).

Table 1 Polysemous Words Tested

Words	Senses	Instances	Resolved
band	group of musicians	19	18
	strip or stripe	4	4
cabinet	administrative organ	24	22
	shelf	17	16
court	judicial	163	158
	area for ball game	9	8
crane	machine	16	16
	bird	21	18
palm	tree	11	11
	hand	52	42
plant	living thing	86	81
	factory	18	15
sentence	group of words	25	22
	punishment	67	59
slug	bullet	6	4
	animal	16	16
tank	combat vehicle	12	12
	water-filled place	10	8
trial	action of judging	89	82
	test	17	15

Table 2 shows the disambiguation result (success rate). Experiment 1, Experiment 2, and Experiment 3 are for $l=10$, 35, and 50, respectively.

Table 2 Disambiguation Result(%)

	band	cabinet	court	crane	palm	plant	sentence	slug	tank	trial	Average
Experiment 1	87.0	82.9	91.9	91.9	87.3	92.3	87.0	77.3	95.5	89.6	89.4
Experiment 2	95.7	92.7	96.5	91.9	84.1	92.3	88.0	90.9	90.9	91.5	91.9
Experiment 3	95.7	97.6	95.9	97.3	84.1	91.3	80.4	72.7	90.9	96.2	91.3

As is seen in Table 2, the overall average success rates are 89.4%, 91.9%, and 91.3%, respectively, for Experiment 1, Experiment 2, and Experiment 3. The best overall result is got when $l = 35$ (Experiment 2). The best single result, 97.6%, is got for *cabinet* when $l = 50$ and the worst is for *slug* when $l = 50$.

6 Evaluation

The upper-bound for the success rate in our experiment is supposed to be 100%, which is that of human beings. The lower-bound or the base-line success rate is 80.5%, which is the

sum of bigger instances in each word divided by the total instances, i.e., 549/682. Obviously, our result is much better than that of the base-line.

Methodologically, our approach is similar to Dagan, Marcus, and Markovitch (Dagan, Marcus and Markovitch 1995) and Schütze (Schütze 1998) in that all use some kinds of word similarity measures. Dagan, Marcus, and Markovitch (Dagan, Marcus and Markovitch 1995) used Augmented TWS method to disambiguate the word senses. It relied on the adjacent word only of a polysemous word to disambiguate the word senses and was doomed to fail, at least for noun disambiguation, since not enough disambiguating information would be provided by the adjacent word alone. In this paper, we considered the global context using a window of l words each before and after the ambiguous word in the text, and took more words into account to make use of enough disambiguating information. Schütze (Schütze 1998) used an idea of Word Space that was a high-dimensional, real-valued space. Then he got vectors of huge dimensions, in which many irrelevant information to express the topic involved was put. Here in our paper, we used the k most similar words and got better vectors, capturing effective information for disambiguating word sense.

Experimentally, our result is better than the ones in other studies (e.g., Chen and Chang 1998; Hiro, Wu and Furugori 1996; Li, Szpakowicz and Matwin 1995). But this does not necessarily mean the superiority of our method: no experiments are done under the same settings and they differ in conditions and data used.

Previously, Peng, Takakura, and Furugori (Peng, Takakura and Furugori 2001) experimented a Dagan-like word similarity measure (Dagan, Marcus and Markovitch 1995) for the same data we used in the present paper. The similarity measure used was:

$$sim(w_1, w_2) = \frac{\sum_{w \in Lexicon} \min(I(w, w_1), I(w, w_2))}{\sum_{w \in Lexicon} \max(I(w, w_1), I(w, w_2))} \quad (7)$$

The success rate got from this experiment was 91.5%. Percentage-wise, this result is comparable with ours, but our present method is superior in the following two points.

First and computationally, the equation (2) makes the calculation time less than that of using equation (7) since $\|S\| \ll \|Lexicon\|$.

Second and methodologically, the equation (2) is free from a deficiency seen in equation (7). Let us see this in examples for $\|Lexicon\| = 8$. There would be cases in using equation (7).

$$\begin{aligned} \text{case 1} \quad (a) \quad sim(w_1, w_2) &= \frac{0.27+0.27+0.27+0.27+0.27+0+0+0}{0.32+0.32+0.32+0.32+0.32+0+0+0} \\ (b) \quad sim(w_3, w_4) &= \frac{2.7+2.7+2.7+2.7+2.7+0+0+0}{3.2+3.2+3.2+3.2+3.2+0+0+0} \\ \text{case 2} \quad (a) \quad sim(w_1, w_2) &= \frac{2.7+2.7+0.2+0.2+0.2+0+0+0}{3.2+3.2+0.3+0.3+0.3+0+0+0} \\ (b) \quad sim(w_3, w_4) &= \frac{2.7+2.7+2.7+2.7+2.7+0.2+0.2+0}{3.2+3.2+3.2+3.2+3.2+0.3+0.3+0} \end{aligned}$$

In *case 1*, 2.7 in (b) is far bigger than 0.27 in (a) and the same is true for 3.2 and 0.32. Nevertheless, we get $\text{sim}(w_1, w_2) = \text{sim}(w_3, w_4)$. In *case 2*, the numbers of the first two terms (2.7, 3.2) in (a) are very big, while those of the rest are very small, and the numbers of the first five terms in (b) are very big, while those of the rest are very small. Nevertheless, we get $\text{sim}(w_1, w_2) \approx \text{sim}(w_3, w_4)$.

As is clear already, these problems are eliminated in equation (2). But then a question remains that why our result is not much better than that of using equation (7). An answer to this question is that $\text{sim}(w_1, w_2)$ and $\text{sim}(w_1, w_3)$ share the mutual information $I(w, w_1)$ in equation (7), so unfavorable effect of *cases 1* and *2* is there not in full force. Another is that the noise effect of *cases 1* and *2* is negligible as we used the small value of k ($=6$) in both experiments. We are sure that the difference in performances between equations (2) and (7) would become clearer as the value of k becomes bigger.

7 Conclusion

We proposed a method to calculate the similarity of two words. Using the similarity metric in equation (2) we designed an approach for WSD that uses the context-similarity vector for resolving ambiguities of polysemous words in text.

Our method is intuitional and its performance is in an acceptable level. However, to improve the performance, it may be worthwhile to consider such points as:

- Dynamic adjustment of parameters d and t for extracting the topic effectively from the context.
- Inversion of the distributional matrix to distributional image in the two dimensional coordinates, and the determination of the sense according to distributional character of the image.

In our method and many others, context is the means to identify the meaning of a polysemous word. But some word senses are nonspecific to topics and a word with such a sense can appear freely in many domains of discourse. It is desirable to give a thought to deal with the meaning of polysemous words used in the nontopical sense (Leacock, Chodorow and Miller 1998). However, within a restriction, we contend that our method is theoretically sound and better in generality and simplicity.

Reference

Chen, J. N. and Chang, J. S. (1998). "A Concept-based Adaptive Approach to Word Sense Dis-

- ambiguation.” In *Proceedings of COLING-ACL '98*, 237-243, Montreal, Quebec, Canada.
- Church, K. and Hanks, P. (1990). “Word Association Norms, Mutual Information and Lexicography.” *Computational Linguistics*, 16, 22-29.
- Dagan, I. and Itai, A. (1994). “Word Sense Disambiguation Using a Second Language Monolingual Corpus.” *Computational Linguistics*, 20(4), 563-595.
- Dagan, I. Marcus, S., and Markovitch, S. (1995). “Contextual Word Similarity and Estimation from Sparse Data.” *Computer Speech and Language*, 9, 123-152.
- EDR. (1998). <http://www.ijnet.or.jp/edr/index.html>
- Gale, W., Church, K. and Yarowsky, D. (1992). “Work on Statistical Methods for Word Sense Disambiguation.” In *Working Notes, AAAI Fall Symposium Series, Probabilistic Approaches to Natural Language*, 54-60.
- Guthrie, J. L., Guthrie, Y. W. and Aidinejad, H. (1991). “Subject-dependent Co-occurrence and Word Sense Disambiguation.” In *Proceedings of the 29th Meeting of ACL*, 146-152, Berkeley: Morgan Kaufmann.
- Hiro, K., Wu, H. and Furugori, T. (1996). “Word-Sense Disambiguation with a Corpus-Based Semantic Network.” *Journal of Quantitative Linguistics*, 3, 244-251.
- Hirst, G. (1987). “Semantic Interpretation and the Resolution of Ambiguity.” Cambridge: Cambridge University Press.
- Karov, Y. and Edelman, S. (1998). “Similarity-based Word Sense Disambiguation.” *Computational Linguistics*, 24(1), 41-59.
- Leacock, C., Chodorow, M. and Miller, G. A. (1998). “Using Corpus Statistics and WordNet Relations for Sense Identification.” *Computational Linguistics*, 24(1), 147-165.
- Lesk, M. (1986). “Automatic Sense Disambiguation: How to Tell a Pine Cone from an Ice Cream Cone.” In *Proceedings of the 1986 SIGDOC Conference*, New York: Association for Computing Machinery.
- Li, X., Szpakowicz, S. and Matwin, S. (1995). “A WordNet-based Algorithm for Word Sense Disambiguation.” In *Proceedings of IJCAI-95*, 1368-1374, Montreal: Morgan Kaufmann.
- Peng, Q., Takakura, S. and Furugori, T. (2001). “Determination of the Meaning of Polysemous Words Using a Word Similarity Measurement.” *SIG Notes, NL-142, Information Processing Society of Japan*, No. 20, 59-66.
- Schütze, H. (1998). “Automatic Word Sense Discrimination.” *Computational Linguistics*, 24(1), 97-123.
- Small, S. and Rieger, C. (1982). “Parsing and Comprehending with Word Experts (A Theory and its Realization).” In W. Lehnert and M. Ringle (Eds.), *Strategies for natural*

- language processing*, 89-147, Hillsdale NJ: Lawrence Erlbaum.
- Text1. (1998). <http://gamp.c.u-tokyo.ac.jp/archive/textdb.htm>
- Text2. (1998). <http://www.infomotions.com/etexts/>
- Walker, D. (1987). "Knowledge Resource Tools for Accessing Large Text Files." In S. Nirembe. (Ed.), *Machine Translation: Theoretical and Methodological issues*, 247-261, Cambridge: Cambridge University Press.
- WordNet. (2001). <http://www.cogsci.princeton.edu/~wn/>
- Yarowsky, D. (1992). "Word-Sense Disambiguation Using Statistical Models of Roget's Categories Trained on Large Corpora." In *Proceedings of COLING-92*, 454-460, Nantes:ICCL.
- Yarowsky, D. (1993). "One Sense per Collocation." In *Proceedings of ARPA Human Language Technology Workshop*, 266-271, Princeton, NJ.
- Yarowsky, D. (1994a). "Decision Lists for Lexical Ambiguity Resolution: Application to Accent Restoration in Spanish and French." In *Proceedings of the 32nd Annual Meeting*, 88-95, Las Cruces, NM. Association for Computational Linguistics.
- Yarowsky, D. (1994b). "A Comparison of Corpus-based Techniques for Restoring Accents in Spanish and French Text." In *Proceedings of the 2nd Annual Workshop on Very Large Text Corpora*, 19-32, Las Cruces, NM.

Qujiang Peng: Qujiang Peng received a B.S. degree in Mathematics and Computer Science from Chongqing Normal University, China, in 1994 and a M.S. degree in Applied Mathematics from Chongqing University, China, in 1997. He is a Ph.D candidate in Computer Science at the University of Electro-Communications, Tokyo, Japan.

Haodong Wu: Haodong Wu received a B.S. and a M.S. in Computer Science from Chongqing University, China, in 1983 and 1986. He got a Ph.D. in Computer Science from the University of Electro-Communications, Tokyo, Japan, in 1997. He is an associate professor at Dokkyo University. His current research interests include Computational Linguistics, Information Retrieval, and Artificial Intelligence. He is a member of the Information Processing Society of Japan, the Association for Natural Language Processing of Japan, and the Association for Computational Linguistics.

Teiji Furugori: Teiji Furugori is a professor in the Department of Computer Science at the University of Electro-Communications, Tokyo, Japan. He has a Ph.D in Computer Science from SUNY at Buffalo. His research interests

include Computational Linguistics, Natural Language Processing, and Cognitive Science. He is a member of ACM, the Information Processing Society of Japan, the Institute of Electronics, Information and Communication Engineers of Japan, and the Association for Natural Language Processing of Japan.

(Received August 10, 2001)

(Accepted January 10, 2002)