CO-OCCURRENCE VECTORS FROM CORPORA VS. DISTANCE VECTORS FROM DICTIONARIES

Yoshiki Niwa and Yoshihiko Nitta
Advanced Research Laboratory, Hitachi, Ltd.
Hatoyama, Saitama 350-03, Japan
{niwa2, nitta}@harl.hitachi.co.jp

Abstract

A comparison was made of vectors derived by using ordinary co-occurrence statistics from large text corpora and of vectors derived by measuring the interword distances in dictionary definitions. The precision of word sense disambiguation by using co-occurrence vectors from the 1987 Wall Street Journal (20M total words) was higher than that by using distance vectors from the Collins English Dictionary (60K head words + 1.6M definition words). However, other experimental results suggest that distance vectors contain some different semantic information from co-occurrence vectors.

1 Introduction

Word vectors reflecting word meanings are expected to enable numerical approaches to semantics. Some early attempts at vector representation in psycholinguistics were the ***semantic differential*** approach (Osgood et al. 1957) and the ***associative distribution*** approach (Deese 1962). However, they were derived manually through psychological experiments. An early attempt at automation was made by Wilks *et al.* (1990) using co-occurrence statistics. Since then, there have been some promising results from using co-occurrence vectors, such as word sense disambiguation (Schütze 1993), and word clustering (Pereira et al. 1993).

However, using the co-occurrence statistics requires a huge corpus that covers even most rare words. We recently developed word vectors that are derived from an ordinary dictionary by measuring the interword distances in the word definitions (Niwa and Nitta 1993). This method, by its nature, has no problem handling rare words. In this paper we examine the usefulness of these ***distance vectors*** as semantic representations by comparing them with co-occurrence vectors.

2 Distance Vectors

A reference network of the words in a dictionary (Fig. 1) is used to measure the distance between words. The network is a graph that shows which words are used in the definition of each word (Nitta 1988). The network shown in Fig. 1 is for a very small portion of the reference network for the Collins English Dictionary (1979 edition) in the CD-ROM I (Liberman 1991), with 60K head words + 1.6M definition words.

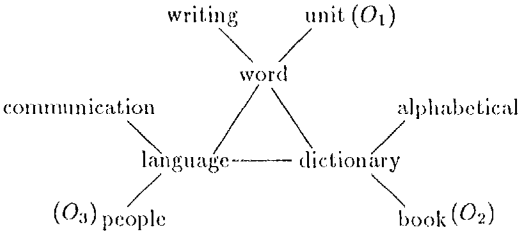

Fig. 1 Portion of a reference network.

For example, the definition for ***dictionary*** is "a book in which the words of a language are listed alphabetically ... ." The word ***dictionary*** is thus linked to the words *book*, *word*, *language*, and *alphabetical*.

A word vector is defined as the list of distances from a word to a certain set of selected words, which we call *origins*. The words in Fig. 1 marked with O_i (*unit*, *book*, and *people*) are assumed to be origin words. In principle, origin words can be freely chosen. In our experiments we used middle frequency words: the 51st to 1050th most frequent words in the reference Collins English Dictionary (CED).

The distance vector for *dictionary* is derived as follows:

$$\textit{dictionary} \Rightarrow \begin{pmatrix} 2 \\ 1 \\ 2 \end{pmatrix} \begin{array}{l} \cdots \text{distance } (\textit{dict.}, O_1) \\ \cdots \text{distance } (\textit{dict.}, O_2) \\ \cdots \text{distance } (\textit{dict.}, O_3) \end{array}$$

The i-th element is the distance (the length of the shortest path) between *dictionary* and the i-th origin, O_i. To begin, we assume every link has a constant length of 1. The actual definition for link length will be given later.

If word A is used in the definition of word B, these words are expected to be strongly related. This is the basis of our hypothesis that the distances in the reference network reflect the associative distances between words (Nitta 1993).

Use of Reference Networks Reference networks have been successfully used as neural networks (by Véronis and Ide (1990) for word sense disambiguation) and as fields for artificial association, such as spreading activation (by Kojima and Furugori (1993) for context-coherence measurement). The distance vector of a word can be considered to be a list of the activation strengths at the origin nodes when the word node is activated. Therefore, distance vectors can be expected to convey almost the same information as the entire network, and clearly they are much easier to handle.

Dependence on Dictionaries As a semantic representation of words, distance vectors are expected to depend very weakly on the particular source dictionary. We compared two sets of distance vectors, one from LDOCE (Procter 1978) and the other from COBUILD (Sinclair 1987), and verified that their difference is at least smaller than the difference of the word definitions themselves (Niwa and Nitta 1993).

We will now describe some technical details about the derivation of distance vectors.

Link Length Distance measurement in a reference network depends on the definition of link length. Previously, we assumed for simplicity that every link has a constant length. However, this simple definition seems unnatural because it does not reflect word frequency. Because a path through low-frequency words (rare words) implies a strong relation, it should be measured as a shorter path. Therefore, we use the following definition of link length, which takes account of word frequency.

$$\mathrm{length}\ (\mathrm{W}_1, \mathrm{W}_2) \underset{def}{=} -\log\left(\frac{\mathrm{n}^2}{\mathrm{N}_1 \cdot \mathrm{N}_2}\right)$$

This shows the length of the links between words $\mathrm{W}_i (i = 1, 2)$ in Fig. 2, where N_i denotes the total number of links from and to W_i and n denotes the number of direct links between these two words.

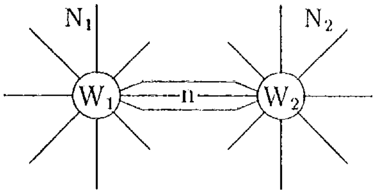

Fig. 2 Links between two words.

Normalization Distance vectors are normalized by first changing each coordinate into its deviation in the coordinate:

$$v = (v_i) \quad \longrightarrow \quad v' = \left(\frac{v_i - a_i}{\sigma_i}\right),$$

where a_i and σ_i are the average and the standard deviation of the distances from the i-th origin. Next, each coordinate is changed into its deviation in the vector:

$$v' = (v'_i) \quad \longrightarrow \quad \overline{v} = \left(\frac{v'_i - \overline{v'}}{\sigma'}\right),$$

where $\overline{v'}$ and σ' are the average and the standard deviation of v'_i $(i = 1, ...)$.

3 Co-occurrence Vectors

We use ordinary co-occurrence statistics and measure the co-occurrence likelihood between two words, X and Y, by the mutual information estimate (Church and Hanks 1989):

$$\mathrm{I}(\mathrm{X}, \mathrm{Y}) = \log^{+} \frac{\mathrm{P}(\mathrm{X} \mid \mathrm{Y})}{\mathrm{P}(\mathrm{X})},$$

where P(X) is the occurrence density of word X in a whole corpus, and the conditional probability P(X | Y) is the density of X in a neighborhood of word Y. Here the neighborhood is defined as 50 words before or after any appearance of word Y. (There is a variety of *neighborhood* definitions such as "100 surrounding words" (Yarowsky 1992) and "within a distance of no more than 3 words ignoring function words" (Dagan et al. 1993).)

The logarithm with '+' is defined to be 0 for an argument less than 1. Negative estimates were neglected because they are mostly accidental except when X and Y are frequent enough (Church and Hanks 1989).

A co-occurence vector of a word is defined as the list of co-occurrence likelihood of the word with a certain set of origin words. We used the same set of origin words as for the distance vectors.

$$\mathrm{CV}[\mathrm{w}] = \begin{pmatrix} \mathrm{I}(\mathrm{w}, \mathrm{O}_1) \\ \mathrm{I}(\mathrm{w}, \mathrm{O}_2) \\ \vdots \\ \vdots \\ \mathrm{I}(\mathrm{w}, \mathrm{O}_m) \end{pmatrix}$$

Co-occurrence Vector.

When the frequency of X or Y is zero, we can not measure their co-occurence likelihood, and such cases are not exceptional. This sparseness problem is well-known and serious in the co-occurrence statistics. We used as a corpus the 1987 Wall Street Journal in the CD-ROM 1 (1991), which has a total of 20M words. The number of words which appeared at least once was about 50% of the total 62K head words of CED, and the percentage of the *word-origin* pairs which appeared at least once was about 16% of total 62K × 1K (=62M) pairs. When the co-occurrence likelihood can not be measured, the value I(X, Y) was set to 0.

4 Experimental Results

We compared the two vector representations by using them for the following two semantic tasks. The first is word sense disambiguation (WSD) based on the similarity of context vectors; the second is the learning of *positive* or *negative* meanings from example words.

With WSD, the precision by using co-occurrence vectors from a 20M words corpus was higher than by using distance vectors from the CED.

4.1 Word Sense Disambiguation

Word sense disambiguation is a serious semantic problem. A variety of approaches have been proposed for solving it. For example, Véronis and Ide (1990) used reference networks as neural networks, Hearst (1991) used (shallow) syntactic similarity between contexts, Cowie *et al.* (1992) used simulated annealing for quick parallel disambiguation, and Yarowsky (1992) used co-occurrence statistics between words and thesaurus categories.

Our disambiguation method is based on the similarity of context vectors, which was originated by Wilks *et al.* (1990). In this method, a context vector is the sum of its constituent word vectors (except the target word itself). That is, the context vector for context,

$$C : \ldots w_{-N} \ldots w_{-1}\ w\ w_1 \ldots w_{N'} \ldots ,$$

is

$$V(C) = \sum_{i=-N}^{N'} V(w_i).$$

The similarity of contexts is measured by the angle of their vectors (or actually the inner product of their normalized vectors).

$$\mathrm{sim}(C_1, C_2) = \frac{V(C_1)}{|V(C_1)|} \cdot \frac{V(C_2)}{|V(C_2)|} .$$

Let word w have senses $s_1, s_2, \ldots, s_m$, and each sense have the following context examples.

Sense	Context Examples
s_1	$C_{11}, C_{12}, \ldots C_{1n_1}$
s_2	$C_{21}, C_{22}, \ldots C_{2n_2}$
$\vdots$	$\vdots$
s_m	$C_{m1}, C_{m2}, \ldots C_{mn_m}$

We infer that the sense of word w in an arbitrary context C is s_i if for some j the similarity, $\mathrm{sim}(C, C_{ij})$, is maximum among all the context examples.

Another possible way to infer the sense is to choose sense s_i such that the average of $\mathrm{sim}(C, C_{ij})$ over $j = 1, 2, \ldots, n_i$ is maximum. We selected the first method because a peculiarly similar example is more important than the average similarity.

Figure 3 (next page) shows the disambiguation precision for 9 words. For each word, we selected two senses shown over each graph. These senses were chosen because they are clearly different and we could collect sufficient number (more than 20) of context examples. The names of senses were chosen from the category names in Roget's International Thesaurus, except *organ*'s.

The results using distance vectors are shown by dots (• • •), and using co-occurrence vectors from the 1987 WSJ (20M words) by circles (○ ○ ○).

A context size (x-axis) of, for example, 10 means 10 words before the target word and 10 words after the target word. We used 20 examples per sense; they were taken from the 1988 WSJ. The test contexts were from the 1987 WSJ: The number of test contexts varies from word to word (100 to 1000). The precision is the simple average of the respective precisions for the two senses.

The results of Fig. 3 show that the precision by using co-occurrence vectors are higher than that by using distance vectors except two cases, *interest* and *customs*. And we have not yet found a case where the distance vectors give higher precision. Therefore we conclude that co-occurrence vectors are advantageous over distance vectors to WSD based on the context similarity.

The sparseness problem for co-occurrence vectors is not serious in this case because each context consists of plural words.

4.2 Learning of *positive-or-negative*

Another experiment using the same two vector representations was done to measure the learning of *positive* or *negative* meanings. Figure 4 shows the changes in the precision (the percentage of agreement with the authors' combined judgement). The x-axis indicates the number of example words for each *positive* or *negative* pair. Judgement was again done by using the nearest example. The example and test words are shown in Tables 1 and 2, respectively.

In this case, the distance vectors were advantageous. The precision by using distance vectors increased to about 80% and then leveled off, while the precision by using co-occurrence vectors stayed around 60%. We can therefore conclude that the property of *positive-or-negative* is reflected in distance vectors more strongly than in co-occurrence vectors. The sparseness problem is supposed to be a major factor in this case.

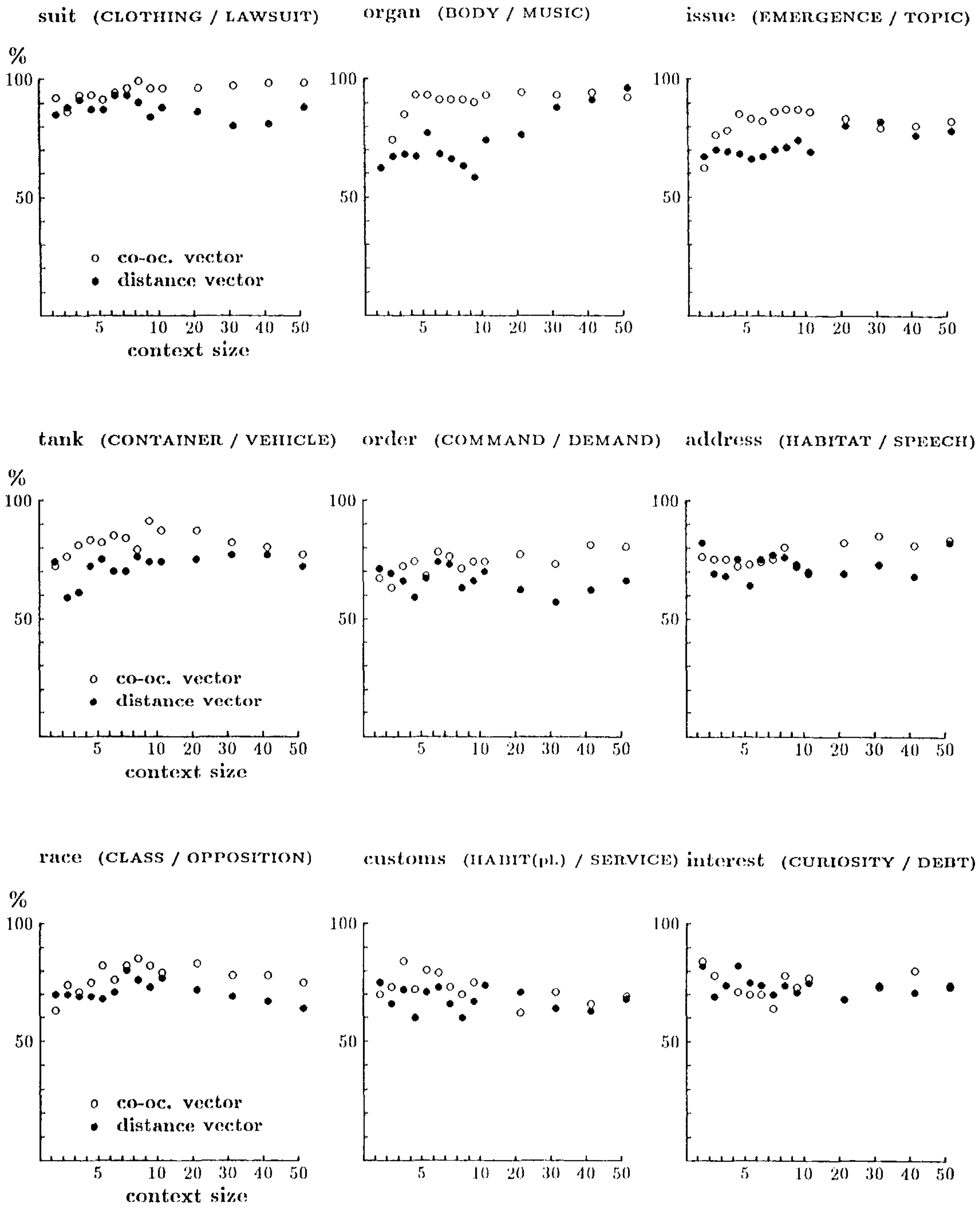

Fig. 3 Disambiguation of 9 words by using co-occurrence vectors(ooo) and by using distance vectors (•••). (The number of examples is 10 for each sense.)

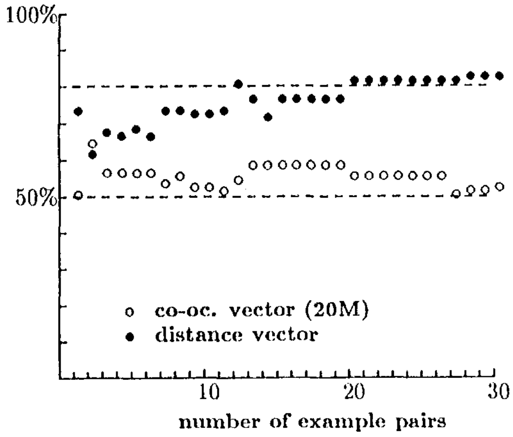

Fig. 4 Learning of *positive-or-negative.*

Table 1 Example pairs.

	positive	*negative*		*positive*	*negative*
1	true	false	16	properly	crime
2	new	wrong	17	succeed	die
3	better	disease	18	worth	violent
4	clear	angry	19	friendly	hurt
5	pleasure	noise	20	useful	punishment
6	correct	pain	21	success	poor
7	pleasant	lose	22	interesting	badly
8	suitable	destroy	23	active	fail
9	clean	dangerous	24	polite	suffering
10	advantage	harm	25	win	enemy
11	love	kill	26	improve	rude
12	best	fear	27	favour	danger
13	successful	war	28	development	anger
14	attractive	ill	29	happy	waste
15	powerful	foolish	30	praise	doubt

Table 2 Test words.

positive (20 words)

balanced	elaborate	elation	eligible	enjoy
fluent	honorary	honourable	hopeful	hopefully
influential	interested	legible	lustre	normal
recreation	replete	resilient	restorative	sincere

negative (30 words)

confusion	cuckold	dally	damnation	dull
ferocious	flaw	hesitate	hostage	huddle
inattentive	liverish	lowly	mock	neglect
queer	rape	ridiculous	savage	scanty
sceptical	schizophrenia	scoff	scruffy	shipwreck
superstition	sycophant	trouble	wicked	worthless

4.3 Supplementary Data

In the experiments discussed above, the corpus size for co-occurrence vectors was set to 20M words ('87 WSJ) and the vector dimension for both co-occurrence and distance vectors was set to 1000. Here we show some supplementary data that support these parameter settings.

a. Corpus size (for co-occurrence vectors)

Figure 5 shows the change in disambiguation precision as the corpus size for co-occurrence statistics increases from 200 words to 20M words. (The words are *suit*, *issue* and *race*, the context size is 10, and the number of examples per sense is 10.) These three graphs level off after around 1M words. Therefore, a corpus size of 20M words is not too small.

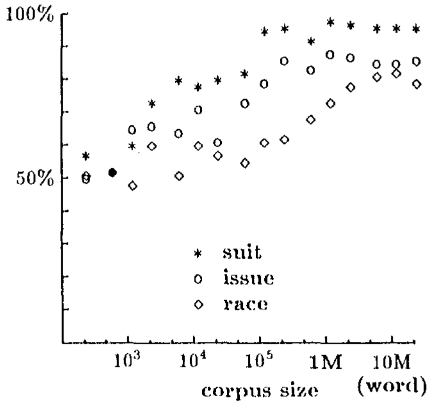

Fig. 5 Dependence of the disambiguation precision on the corpus size for co-occurrence vectors.
context size: 10,
number of examples: 10/sense,
vector dimension: 1000.

b. Vector Dimension

Figure 6 (next page) shows the dependence of disambiguation precision on the vector dimension for (i) co-occurrence and (ii) distance vectors. As for co-occurrence vectors, the precision levels off near a dimension of 100. Therefore, a dimension size of 1000 is sufficient or even redundant. However, in the distance vector's case, it is not clear whether the precision is leveling or still increasing around 1000 dimension.

5 Conclusion

- A comparison was made of co-occurrence vectors from large text corpora and of distance vectors from dictionary definitions.
- For the word sense disambiguation based on the context similarity, co-occurrence vectors from the 1987 Wall Street Journal (20M total words) was advantageous over distance vectors from the Collins English Dictionary (60K head words + 1.6M definition words).
- For learning *positive* or *negative* meanings from example words, distance vectors gave remarkably higher precision than co-occurrence vectors. This suggests, though further investigation is required, that distance vectors contain some different semantic information from co-occurrence vectors.

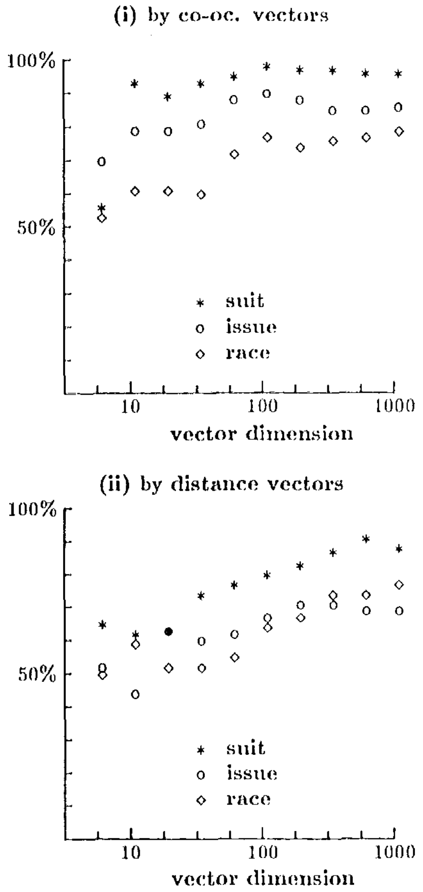

Fig. 6 Dependence on vector dimension for (i) co-occurrence vectors and (ii) distance vectors. context size: 10, examples: 10/sense, corpus size for co-oc. vectors: 20M word.

References

Kenneth W. Church and Patrick Hanks. 1989. Word association norms, mutual information, and lexicography. In *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics*, pages 76–83, Vancouver, Canada.

Jim Cowie, Joe Guthrie, and Louise Guthrie. 1992. Lexical disambiguation using simulated annealing. In *Proceedings of COLING-92*, pages 359–365, Nantes, France.

Ido Dagan, Shaul Marcus, and Shaul Markovitch. 1993. Contextual word similarity and estimation from sparse data. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, pages 164–171, Columbus, Ohio.

James Deese. 1962. On the structure of associative meaning. *Psychological Review*, 69(3):161–175.

Marti A. Hearst. 1991. Noun homograph disambiguation using local context in large text corpora. In *Proceedings of the 7th Annual Conference of the University of Waterloo Center for the New OED and Text Research*, pages 1–22, Oxford.

Hideki Kozima and Teiji Furugori. 1993. Similarity between words computed by spreading activation on an english dictionary. In *Proceedings of EACL-93*, pages 232–239, Utrecht, the Netherlands.

Mark Liberman, editor. 1991. *CD-ROM I.* Association for Computational Linguistics Data Collection Initiative, University of Pennsylvania.

Yoshihiko Nitta. 1988. The referential structure of the word definitions in ordinary dictionaries. In *Proceedings of the Workshop on the Aspects of Lexicon for Natural Language Processing, LNL88-8, JSSST*, pages 1–21, Fukuoka University, Japan. (in Japanese).

Yoshihiko Nitta. 1993. Referential structure – a mechanism for giving word-definition in ordinary lexicons. In C. Lee and B. Kang, editors, *Language, Information and Computation*, pages 99–110. Thachaksa, Seoul.

Yoshiki Niwa and Yoshihiko Nitta. 1993. Distance vector representation of words, derived from reference networks in ordinary dictionaries. MCCS 93-253, Computing Research Laboratory, New Mexico State University, Las Cruces.

C. E. Osgood, G. F. Such, and P. H. Tannenbaum. 1957. *The Measurement of Meaning.* University of Illinois Press, Urbana.

Fernando Pereira, 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, Columbus, Ohio.

Paul Procter, editor. 1978. *Longman Dictionary of Contemporary English (LDOCE).* Longman, Harlow, Essex, first edition.

Hinrich Schütze. 1993. Word space. In J. D. Cowan S. J. Hanson and C. L. Giles, editors, *Advances in Neural Information Processing Systems*, pages 895–902. Morgan Kaufmann, San Mateo, California.

John Sinclair, editor. 1987. *Collins COBUILD English Language Dictionary.* Collins and the University of Birmingham, London.

Jean Véronis and Nancy M. Ide. 1990. Word sense disambiguation with very large neural networks extracted from machine readable dictionaries. In *Proceedings of COLING-90*, pages 389–394, Helsinki.

Yorick Wilks, Dan Fass, Cheng ming Guo, James E. McDonald, Tony Plate, and Brian M. Slator. 1990. Providing machine tractable dictionary tools. *Machine Translation*, 5(2):99–154.

David Yarowsky. 1992. Word-sense disambiguation using statistical models of roget's categories trained on large corpora. In *Proceedings of COLING-92*, pages 454–460, Nantes, France.