Chapter 11

Using Pathfinder to Extract Semantic Information from Text

James E. McDonald, Tony A. Plate, and Roger W. Schvaneveldt

How might the semantic information contained in existing textual material, such as dictionaries, be made more "tractable" for use by machines? In this chapter we discuss a technique for extracting relatedness information from text along with some potential uses for such information. The method is based on frequency of co-occurrence, that is, the number of times pairs of words occur together in selected units of text. We hypothesize that frequency of co-occurrence provides a reasonable basis for estimating relatedness among the objects, events, situations, states, and so forth, that "words" refer to, particularly for certain applications, and that it offers several advantages over the use of human judges to establish such estimates. We have been encouraged by the results of applying this method to the analysis of the *Longman Dictionary of Contemporary English* (LDOCE), and this effort has allowed us to identify several questions in need of additional research as well.

Although the focus of the present work is LDOCE, we believe that our method may be useful in other applications as well.[1] In the following sections we discuss (1) our objectives for this work, (2) co-occurrence data and relatedness functions derived from them, (3) the use of Pathfinder networks to simplify the representation of co-occurrence data, (4) some experiments aimed at validating the use of frequency of co-occurrence data to estimate the strength of relationships among words, and (5) two approaches to lexical sense selection, one which uses co-occurrence data directly and another which uses Pathfinder networks derived from co-occurrence data.

Our objectives for this work are primarily practical, although there are certainly theoretical implications for computational linguistics and cognitive psychology as well. Our immediate goal is to produce a modified version of LDOCE, one in which the words in the definitions have appropriate sense tags. The method will need to be refined, however, since our long-term objective is to make word-sense distinctions for unconstrained natural-language (general lexical sense selection). This is not meant to suggest that the statistical approach we propose is capable of solving all the problems of natural language understanding—or even lexical sense selection. Rather, we are investigating how a subsystem using co-occurrence can be built so that it will be useful as part of a natural-language understanding system.

[1] We have, for example, applied the same method to the UNIX online documentation system (the *man* system) as part of an effort to build a hypertext browser for UNIX.

The Nature of Co-occurrence Data

We assume that words co-occur in sentences because they are related to the idea being expressed by the sentence (the meaning of the sentence). The words are therefore semantically related in the context of the sentence. Our technique for lexical sense selection only relies on this claim being true in the aggregate—which is fortunate because individual sentences can certainly be constructed which violate this assumption. Strictly speaking, we contend that pairs of words co-occur frequently in *collections* of sentences because they are semantically related. In testing this claim we will discuss some experiments in which estimates of relatedness derived using our method are compared with human judgments of relatedness. The sense-selection experiments themselves also serve to evaluate this contention.

Co-occurrence in Text

In text, co-occurrence data record the frequencies of co-occurrence of pairs of words within some textual unit. The textual unit can be a phrase, a sentence, a paragraph, or any other identifiable unit. The co-occurrence data used in the work reported in this chapter were collected using the sense-definition as the textual unit.[2]

> The frequency of co-occurrence of two words is defined as the number of textual units in which some form of both of those words occurs. For two words, x and y, their co-occurrence frequency is designated f_{xy}.

The independent frequencies of occurrence of words in a textual unit are also important and are used in conjunction with frequencies of co-occurrence to calculate the values of various relatedness functions.

> The independent frequency of occurrence of a word is defined as the number of textual units in which some form of that word occurs. The frequency of occurrence of word x is designated f_x.

Extracting Co-occurrence Data from LDOCE

Some of the methods used to convey word meaning in dictionary definitions include giving examples of use in context (*illustrative definitions*), saying something directly about meaning (*descriptive definitions*), or simply providing other words with the same meaning (*synonym definitions*). In fact, all three methods are used in various combinations in different dictionaries. LDOCE, for example, relies primarily on descriptive and illustrative techniques, although cross-references (synonyms and related words) are often provided. However, the objective that "the definitions are always written using simpler terms than the words they describe," expressed in the introduction to LDOCE, limits the extent to which synonymy can be used in defining words (e.g., *copy* is used in the definition of one of the senses of *reproduce,* but not visa versa).

Although co-occurrence statistics could be collected on free text, and might prove useful, a dictionary such as LDOCE offers certain advantages. First, unique to LDOCE, the vocabulary used in defining word senses is limited (the LDOCE-controlled vocabulary contains approximately 2,187 words). A limited vocabulary makes the task of collecting

[2]A sense-definition is considered to be the entire definition of a sense of a word, including any examples. Sense definitions can be easily identified in LDOCE.

and storing frequencies of co-occurrence simpler, allowing the use of conventional computing techniques without requiring excessive resources.[3]

The second advantage of dictionaries for our purposes is that sense definitions provide small, coherent units of text centered around single ideas, natural units of co-occurrence. Other sources of text, such as thesauri and encyclopaedis, also provide coherent units of text, but free text is not as constrained. The point is that free text would not be as efficient at providing relatedness estimates among the words in the controlled vocabulary. This conjecture has been verified to some extent by comparing the co-occurrence information taken from the definitions in LDOCE with that obtained from the example sentences. In this comparison, estimates of relatedness derived from definitions correlated more highly with human subjects' ratings of relatedness than did estimates based on example sentences. This does not rule out the usefulness of free text as a source of relatedness information, but does support the claim that the definitions in LDOCE provide semantically focused units for co-occurrence analysis.

The last advantage of dictionaries over free text is that dictionaries provide definitions for all word senses. It is therefore possible to build a representation for every sense contained in the dictionary, not just those that occur frequently. Although it may be necessary to augment these representations through the use of other dictionaries or perhaps even free text, at least minimal representations can be obtained for even infrequent sense distinctions. Unfortunately, LDOCE doesn't contain very many words (about 1.2 million in total). Therefore, for many words in the controlled vocabulary the frequency of occurrence is quite low (280 occur 30 or fewer times). More importantly, not all of the senses of words are used in defining other words, limiting the accuracy of relatedness estimates that can be obtained from co-occurrence data.

Related Work

The automatic extraction of useful information from text has been a long-standing goal of several investigators. These efforts have ranged from work in artificial intelligence aimed at text understanding to systems for automatically indexing collections of documents. In the field of information retrieval, much of the work has focused on methods for determining the content of documents by examining the individual words contained in titles, abstracts, or entire documents (Belkin & Croft, 1987; Brooks, 1987; Dumais, 1988; Salton, 1986). Most of the attempts to develop automatic retrieval systems have relied, in various ways, on comparing the words that occur in queries and the words that occur in documents.

Many of the failings of indexing using individual words can be attributed to the complexities of the relations between words and meanings. Homography, polysemy, and synonomy all contribute to uncertainty about the similarities and differences in the meaning of words occurring in different documents. In various contexts, the same words can mean quite different things, and different words can have quite similar meanings. If the appropriate senses of homographs and polysemous words in a text could be determined, more precise comparisons could be made between the meanings occurring in different texts. One of the goals of our own work has been to develop methods of identifying the sense of a word in text. Others have had similar goals.

Lesk (1986) attempted to identify word senses by comparing the words in a sentence containing a target word with the words occurring in the sense definitions of the target word. He reports some success with the method (50-70% correct sense selections), and he

[3]For example, although the array of frequencies of co-occurrence for LDOCE requires approximately 4.7 megabytes of storage, it takes less than an hour to build on a minicomputer.

speculates that dictionaries with longer definitions would improve the performance. As we show later, better performance in identifying word senses is achieved by expanding the set of words to be considered by adding other words associated with the words in the context sentence. To the extent that expanding word sets produces the same result as longer definitions, this finding offers some support for Lesk's conjecture. Lesk (1987) uses a similar system based on overlap of words in dictionary definitions to identify words related to words in an information query. The use of this technique to identify related or associated words is very similar to our efforts to find relatedness estimates from the co-occurrence of words in LDOCE. Our work uses co-occurrences of words throughout the dictionary, in contrast to Lesk's method which looks for co-occurrences within the definitions of the words in question. We also go beyond the relatedness estimates derived from co-occurrence to create a network of interrelated words. The network can then be used to find sets of words related to any particular word.

The use of latent semantic indexing for information retrieval also attempts to identify a structure that represents the pairwise relatedness between words (Dumais, 1988; Dumais, Furnas, Landauer, Deerwester, & Harshman, 1988). This technique attempts to extract orthogonal factors from a large matrix of associations between terms and text objects. These factors can be used to define a semantic space in which both queries and text objects can be placed. Retrieval of text objects can then be based on their proximity to a query in this space. Fowler and Dearholt (Chapter 12, this volume) report a related approach to text and query representation using Pathfinder networks. Both of these efforts are attempt to go beyond matching words to uncovering some of the semantic structure that accompanies the use of words in documents.

Cohen and Kjeldsen (1987) developed a system to match grant proposals with funding agencies based on constrained spreading activation in semantic networks. A network of research topics is connected both to funding agencies and to a research proposal. Activation over these pathways identifies funding agencies who are likely to be interested in the proposal.

One approach to expanding word sets is to identify synonyms via a thesaurus (Furnas, Landauer, Gomez, & Dumais, 1987; Sparck Jones, 1986). Sparck Jones has conducted extensive studies of thesaurus generation.

Sparck Jones (1986) was concerned with finding thesaurus-like groups of words that could be used to resolve lexical ambiguity. She considered her work to be primarily practical, aimed ultimately at machine translation and discourse analysis. As a consequence, one of her objectives was the construction of a machine-tractable dictionary consisting of synonym definitions and organized according to "similarity."

At the core of the Sparck Jones approach is a method for precisely defining synonymy. This method assumes a model of language in which there are multiple signs for word senses and, possibly, multiple "similar" word senses for the same sign.[4] Unfortunately, it isn't clear from the description provided by Sparck Jones how much of the task of constructing the dictionary is to be accomplished by humans and how much by machine.[5] It

[4]Although clearly aware of homography and its impact, Sparck Jones seems to assume that it can be safely ignored. The model of language that she claims represents natural language, her Model 4, assumes that if two or more word uses have the same sign they are similar in meaning.

[5]Sparck Jones seems to believe that many of the definitions in dictionaries consist of synonyms or sets of synonyms, particularly in the *Oxford English Dictionary* (OED). However, in a footnote (Section 6.2) she states that "the thesaurus information in the OED is so inaccessible and so unsystematic that it can be said that it is hardly there at all" (1986, p. 259-260). Whichever is the case for the OED, definition by synonymy is not typically used in LDOCE.

seems likely, however, that considerable human participation would be required since only limited, hand-constructed examples were provided by Sparck Jones.[6]

In her thesis, Sparck Jones argued that it was essential to use a well-defined *linguistic* relationship as the basis for the analysis of natural language by machine. She selected synonymy for theoretical reasons, rejecting several alternatives, such as association and collocation, and went about operationally defining synonymy. She defined a "row" as a set of words (word signs) which can replace each other in a particular sentence without changing its meaning. Thus, the word signs in a row are, very precisely, synonyms and represent a particular word use (sense). Rows, each of which embodies the word use of the word-signs it contains, are *similar* to the extent that they have word signs in common. Although based on synonymy, the nature of the weaker "similarity" relation established by the Sparck Jones method is less clear, but it doesn't appear to be synonymy.

Sparck Jones believed that synonymy is fundamental to language. She rejected methods for establishing semantic relations based on co-occurrence on the grounds that the results may be due to extralinguistic factors. We agree that words can co-occur for many reasons, some pragmatic. However, we believe that this fact doesn't rule out the use of associations derived from co-occurrence for lexical ambiguity resolution. It is commonly recognized that natural-language processing systems will need to incorporate world knowledge in order to be successful. Such relationships can be established automatically using our method. Although the objective of being able to precisely label the relationships among word uses appears desirable, it may not be necessary. We define relatedness in terms of the operations used to establish it, much as synonymy is operationally defined in Sparck Jones' work. In a practical endeavor such as ours, the important questions seem to be "Can it be done?" and "Does it work?"

In spite of the apparent difference, however, there is a rather direct relationship between these two techniques. In her dissertation, Sparck Jones (1986) acknowledged that words can be classified on the basis of common rows, rather than classifying rows on the basis of common words. This is essentially the technique we employ, except that we use sense definitions rather than rows. Sparck Jones went on to speculate that this alternative approach might be more appropriate for machine translation, but that there may be some benefit in using both classification schemes.

Sparck Jones contended that because we don't know how to do machine translation it is difficult or impossible to evaluate the utility of semantic classification. We have taken a different position. We believe that the utility of the semantic classification approach can be assessed prior to completely solving the machine translation problem. Although there might be immediate gains to be had from incorporating syntactic analysis, the semantic and syntactic component are conceptually independent, and they can be evaluated independently.

Explorations of the Co-occurrence Data

We begin this section by emphasizing the magnitude of the task at hand. The co-occurrence data obtained for the LDOCE-controlled vocabulary consists of nearly two-and-a-half

[6]The decision that two words are substitutable in a particular sentence seems to require a very complex linguistic judgment. Such a decision could only be made by a system capable of understanding natural language. One objective of Sparck Jones was to accomplish limited natural-language processing, such as machine translation and discourse analysis. It is impractical to require a machine capable of understanding natural language in order to build one.

million frequencies of co-occurrence (the triangle of a 2187 × 2187 matrix). It is, therefore, impossible to examine these data in raw form. The information must be reduced in some way to be useful. The objective, of course, is to reduce the data while preserving interesting or useful information. In what follows we will discuss two techniques for getting at the important information in the LDOCE co-occurrence data. The first of these uses thresholds and relatedness functions derived from the LDOCE co-occurrence matrix directly. The second technique uses the Pathfinder network scaling algorithm (Schvaneveldt, Dearholt, & Durso, 1988; Schvaneveldt, Durso, & Dearholt, 1989) to determine relatedness and to identify important relationships.

Relatedness Functions

If related words are more likely to occur together than unrelated words, then statistics of co-occurrence provide some indication of relatedness. The problem is to find some function of co-occurrence that reflects the relatedness of pairs of words, that is, a function that will yield the relative strength of relatedness for various pairs of words. We will refer to such functions as *relatedness functions*.

A good relatedness function should have a number of characteristics. Ideally, it should be equally sensitive across the range of independent frequencies. In other words, estimates of relatedness should be independent of the base frequencies of the words involved. This is a particularly difficult characteristic to obtain in practice, and many of the relatedness functions discussed below produce more "accurate" estimates of relatedness when the words have approximately equal independent frequencies than when they differ greatly in frequency. The relatedness function should, of course, also produce valid results. In our evaluations, a good relatedness function will provide a measure that correlates with human judgments of relatedness and one that is successful in selecting appropriate senses of words in sentences.

We examined several relatedness functions with various characteristics (cf. Salton, 1968). Some of the relatedness functions we have considered are shown in Table 1, along with comments about their *bias*, *sensitivity*, and *symmetry*. Bias refers to the extent to which sensitivity varies with the independent frequency of words. Sensitivity refers to the extent to which a measure varies as dependencies in the occurrence of words in a pair vary from chance co-occurrence to maximum possible dependence. Since the maximum possible co-occurrence depends on base frequency, this definition of sensitivity makes sensitivity independent of frequency. Symmetric measures are the same for $f(x,y)$ and $f(y,x)$, whereas asymmetric measures may yield different estimates of relatedness for the two uses of the function.

As an example of the differences produced by applying the various relatedness functions, about 20 of the words most strongly related to *bank* for each of the relatedness functions are shown in Table 2.

Some of these functions include more closed-class words (especially determiners and very common prepositions) in the set of highly related words. Such words seem to provide very little semantic information. The $dcp_{\min}$ and *iou* functions yield the best sets of related words on intuitive grounds. Because the very common closed-class words do not provide much information about the meaning of other words, the most common of these were omitted from the sense selection experiments discussed below.[7]

[7]For the purposes of the experiments described in this section, the following words were omitted from the controlled vocabulary: *a, and, be, for, in, of, or, than, that, the, this, those, to, what, when, where, which, who, with.*

We examined the relatedness functions as well as raw frequency of co-occurrence (*coc*) as measures of relatedness by comparing them to human judgments of relatedness and by attempting to identify the senses of words in sentences using the relatedness functions.

Table 1. Relatedness functions.

Name	*Value*	*Comments*
$cp(x,y)$	$\frac{f_{xy}}{f_y}$ $(= Pr(x \mid y))$	Conditional *p*robability of x given y. Asymmetric. Insensitive and heavily biased for all f_x and f_y, except low, equal values.
$dcp(x,y)$	$Pr(x \mid y) - Pr(x)$	(*d*eviation of *cp*) Asymmetric. More sensitive than *cp* but still biased. An attempt to remove some of the bias of *cp*.
$dcp_{\min}(x,y)$	$\min(dcp(x,y), dcp(y,x))$	Minimum of *dcp* in both directions. Symmetric. Sensitive if f_x and f_y are similar, but results in zero if they are considerably different.
$iou(x,y)$	$Pr(x \text{ and } y \mid x \text{ or } y)$	(*i*ntersection *o*ver *u*nion) Produced by dividing the number of units containing x and y by number of units containing at least one of them. More sensitive than $dcp_{\min}$ when f_x and f_y are different.
$dex(x,y)$	$\frac{f_{xy} - f_x \cdot f_y}{\min(f_x f_y) - f_x \cdot f_y}$	(*d*ependency *ex*traction) Normalizes f_{xy} by mapping it to [0,1] according to its scaled position between its minimum and maximum possible values. Symmetric. Fully sensitive for all f_x and f_y.

Table 2. Words most strongly related to *bank* for each relatedness function.

cp	*a account an and as be by for from have in money of on or river the to which*
dcp	*a account as at be by from have in keep money of on pay river rob the to water*
$dcp_{\min}$	*account cheque criminal earn flood flow lake lend money pay prevent promise rate river rob rock safe sand sum thief*
iou	*account busy cheque criminal earn flood flow interest lake lend money overflow pay river rob safe sand thief wall*
dex	*a account be by cheque clerk dollar in messenger money of overflow participle pay river rob September the to*

Human Judgments

Although there are many questions that might be asked regarding the application of the technique we propose, one of the first seems to be the extent to which the associations derived from text are like ratings of relatedness supplied by human judges. If it can be established that frequency of co-occurrence data, or some transformation of them, correspond to human judgments of relatedness, then we should be able to use this approach for applications that typically require such information. Furthermore, by comparing proximity estimates based on co-occurrence with those obtained from human judges, it should be possible to specify the ways in which these estimates differ and to compensate for such differences if necessary.

The general procedure used in each of the studies in this section was (1) to select a set of LDOCE-controlled vocabulary, (2) obtain judgments of relatedness from human subjects for the selected words, (3) compute estimates of relatedness from the LDOCE co-occurrence matrix, and (4) compare the obtained human judgments and LDOCE-based estimates. The basic measures of correspondence were correlations between relatedness estimates and human judgments.

The human-judgment data for the following comparisons were all obtained using the method of paired comparison. We correlated these data with raw co-occurrence counts (*coc*), average conditional probabilities (*cp*), deviations of conditional probabilities (*dcp*), minimum of *dcp* (dcp_{min}), intersection over union (*iou*), and dependency extraction (*dex*) derived from LDOCE.

The Natural Category Set. For our first comparison we selected the 25 natural category words (e.g., *animal, plant, dog, rose*) for which estimates of relatedness had been obtained from 24 introductory psychology students and 24 biology graduate students (cf., Schvaneveldt, Durso, & Dearholt, 1989). Sixteen of the words used in these rating studies were in the LDOCE-controlled vocabulary. This set of 16 words served as the basis for the following comparisons.

Table 3. Correlations of human judgments of relatedness and estimates of associations derived from LDOCE co-occurrence for 16 primitives.

	A	B	C	D	E	F	G	H
Psychology Students - A		.94	.50	.69	.68	.58	.60	.68
Biology Students - B	.94		.50	.68	.68	.58	.59	.68
LDOCE *coc* - C	.50	.50		.83	.82	.95	.95	.71
LDOCE *cp* - D	.69	.68	.83		1.00	.90	.93	.98
LDOCE *dcp* - E	.68	.68	.82	1.00		.89	.92	.98
LDOCE dcp_{min} - F	.58	.58	.95	.90	.89		.99	.79
LDOCE *iou* - G	.60	.59	.95	.93	.92	.99		.84
LDOCE *dex* - H	.68	.68	.71	.98	.98	.79	.84	

As can be seen from Table 3, the correlations of the LDOCE conditional probabilities with human judgments are quite high, but not as high as the correlation between the two groups of human judges. These results are promising in that we are able to account for a significant amount of the variability in human ratings (48%). The two relatedness measures (*cp* and *dcp*) correlate about the same with human judgments, and for this set of words, the two measures correlate almost perfectly with one another.

The Bank Set. Unlike the previous comparison, the set of words related to *bank* was selected directly from LDOCE. The first step consisted of obtaining all of the words that were associated with *bank* above a relatedness threshold of .01. In turn, sets of words were obtained for each of these words, which resulted in a fairly large set of associated words. The 20 words with the highest number of co-references were selected for the rating experiment. Five Computing Research Laboratory researchers served as subjects in the rating study. The correlational analyses are shown in Tables 4 and 5.

Table 4. Correlations of human judgments of relatedness for the *bank* primitives.

	JM	*TP*	*RS*	*CE*	*JB*
JM		.80	.83	.70	.78
TP	.80		.78	.70	.75
RS	.83	.78		.73	.78
CE	.70	.70	.73		.82
JB	.78	.75	.78	.82	

Table 5. Correlations of relatedness estimates from human judges and LDOCE for the *bank* primitives.

	A	B	C	D	E	F	G
Mean Ratings - A		.48	.66	.65	.60	.64	.61
LDOCE *coc* - B	.48		.73	.71	.68	.76	.66
LDOCE *cp* - C	.66	.73		1.00	.78	.96	.98
LDOCE *dcp* - D	.65	.71	1.00		.79	.86	.98
LDOCE dcp_{min} - E	.60	.68	.78	.79		.98	.64
LDOCE *iou* - F	.64	.76	.86	.86	.98		.73
LDOCE *dex* - G	.61	.66	.98	.98	.64	.73	

The results from these comparisons are again promising. Intersubject correlations ranged from .70 to .83. Correlations between ratings and LDOCE measures are also quite high (except for the direct co-occurrence measure). This set of words produces differences among the different measures derived from LDOCE co-occurrences. Conditional probability is better than frequency of co-occurrence, and nothing is gained from the more complex measures.

Lexical Ambiguity Resolution

Our general method for identifying word senses is relatively straightforward. There are, however, numerous refinements to be considered, some of which will be discussed later in this section. The method requires determining related-word sets for individual words. The related-word sets are defined in different ways, but each set is essentially selected on the basis of some index of relatedness. All words in the controlled vocabulary satisfying some minimum threshold of relatedness with a particular word are included in a related-word set for that word. These related-word sets may also be expanded by including words that are related to the words in the related-word set, and so on. The basic methodology consists of the following steps:

1) A context sentence is selected and a test word is selected from it for sense tagging (lexical sense selection).
2) A context set *c* is formed for the context sentence by combining related-word sets for each word in the context sentence, *except for the test word itself.*
3) Separate definition sets ($d_1 \ldots d_n$ where n is the number of sense definitions for the test word) are formed for each of the sense definitions for the test word using the words in the sense definition, *except for the test word itself.*
4) The proportion of words in each of the definition sets contained in *c* is computed.
5) The sense definition with the largest overlap with words in *c* is judged the winner and the test word is tagged with that sense.

This process is, of course, not as simple as it sounds. We have already discussed the problem of measuring relatedness. There are also various ways to expand the context and definition sets. Identifying the winner can also be done in a variety of ways. These issues are discussed in more detail below.

The Direct Use of Co-occurrence Data

Using the overlap of context sets and definition sets to identify word senses is not simply a "keyword search" approach because the use of related-word sets makes the decision depend on more than the words that a sentence (the context) and the definitions have in common. Often sentences will not share any words with the appropriate definition. For example, the definition of sense 4.1 of *bank* is shown below, followed by an alphabetized list of the base forms of the controlled vocabulary words, excluding the list of ignored words.[8]

> ***bank*$^{4.1}$**: A place in which money is kept and paid out on demand, and where related activities go on. (*activity demand go keep money on out pay place related*)

[8]We use the convention of numbering the M^{th} sense in the N^{th} entry (homograph) for a word as "sense *N.M.*"

An example of the use of sense 4.1 of *bank* is:

> Any of various kinds of bank accounts earning higher interest than a deposit account. (*account any earn high interest kind various*)

Notice that the phrasal context shown above and the sense-entry for *bank* 4.1 have no words in common in the controlled vocabulary. If we consider the words in parentheses to be sets of words, their intersection is empty. This is not at all unusual in LDOCE, given the small number of words used in sense definitions. As a consequence, however, the straightforward technique of looking for the sense-entry which has the maximum intersection with the context doesn't always work well. Our approach to this problem expands the contexts and/or sense-entries to included related words, thereby making the intersection technique more reliable.

In this first series of experiments, we represented expanded contexts and sense-definitions as vectors of weights over all of the LDOCE primitives. Each weight represents the "strength" of the association between a particular word and the context or sense-definition set to which it belongs. Expanded sets were created by adding to each set all of the LDOCE primitives that exceeded a relatedness threshold using one of the relatedness functions. For each expanded set, the weights were the number of words in that set to which each word in the set is related according to the relatedness threshold. These weights were intended to represent the centrality or importance of words in the context or sense-definitions in the sense that the more words in the set that are related to a particular word, the higher its weight. Of course, words that are not in the expanded set have weights of zero. Vectors of weights can be treated as sets by converting non-zero weights to one and zero weights to zero. This operation is expressed as $X > 0$ below, where X is a vector of weights.

Once the context and sense-definitions have been represented as vectors of weights, an estimate of their "similarity" is computed (i.e., the strength of the relationship between the two vectors). All of the functions used measure vector overlap in one way or another. Some of them consider weights, others are based only on set membership.

In the similarity functions, *SUM* sums all of the elements of a vector. The dot-product function "$\cdot$" is the sum of the cross-products. The following similarity functions were used in various experiments.

The commonality or *COM* function treats context and sense-definition vectors as sets.

$$COM(V, W) = \frac{|(V>0) \cap (W>0)|}{|(V>0) \cup (W>0)|}$$

$HIT^{\rightarrow}(V,W)$ counts the "hits" of V in W (i.e., it sums the weights for the words in the intersection of V and W) and divides this value by the sum of the weights in W. The right-pointing arrow is used to indicate that this is an asymmetrical function [i.e., $HIT^{\rightarrow}(V,W)$ is not necessarily equal to $HIT^{\rightarrow}(W,V)$].

$$HIT^{\rightarrow}(V,W) = \frac{(V>0)\cdot W}{SUM(W)}$$

HIT^{x} takes the product of $HIT^{\rightarrow}$ to produce a symmetric result.

$$HIT^{x}(V,W) = HIT^{\rightarrow}(V,W)\, HIT^{\rightarrow}(W,V)$$

Finally, we have found it useful to compute the similarity between two vectors using the normalized dot-product (i.e., the cosine of the angle between the two vectors).

$$NDP(V,W) = \frac{(V \cdot W)}{\sqrt{V \cdot V + W \cdot W}}$$

Using the general lexical ambiguity resolution procedure already described, we attempted to select the correct sense of *bank* for the 197 sentences containing the word *bank* in LDOCE. The test sentences were first manually sense-tagged by the authors using the sense distinctions made in LDOCE for *bank*. This was not always a simple task because, in the judgment of the authors, some of the usages of *bank* cannot be classified using the sense distinctions for *bank* in LDOCE. More generally, there is some question as to whether or not all of the sense distinctions made in LDOCE are legitimate or, conversely, whether particular sense distinctions are missing. Nevertheless, the automatic method was judged correct if it chose the same sense as that selected by the authors beforehand.

The word *bank* was selected as a test case for a number of reasons. First, it has a "moderate" number of sense distinctions (13), at least as far as words in LDOCE go, yet the senses of *bank* can be easily divided into larger groups. The two main (homographic) sense "groups" contain *financial* senses and *earth* or *river* senses, respectively. These two groups account for 7 of the 13 senses, and, more importantly, nearly all of the usages of *bank* in LDOCE. Some of the finer sense distinctions within these two groups are semantic, whereas others are syntactic. For example, two of the three *financial* senses of *bank* are the transitive and intransitive verbal forms. Because our method does not directly consider syntactic information, we did not expect it to be able to correctly discriminate these uses. We also suspected that the method would have difficulty discriminating among the *earth* senses of *bank*, which differ by fine semantic distinctions. These considerations led us to construct groups of senses that contained gross rather than fine semantic distinctions. For this purpose, we assigned the 13 senses to 6 *sense-groups*, and the ability of the method to assign occurrences of *bank* to these larger groups was also assessed.[9]

Identifying the correct sense of *bank* proved to be a difficult task. In only 38 of 420 experiments was *bank* correctly sense-tagged 35% or more of the time. However, the probability of correctly tagging a particular usage of *bank* by chance is only 7.7%. As expected, selecting the correct sense-group was a far easier task. In 120 of the experiments, *bank* was assigned to the correct sense group 85% or more of the time (the probability is only 17% by chance). The experiments yielding the best performance are summarized in Table 6. The results of using the NDP similarity function using only the words in the context sentence and in the definitions are included for comparison.

[9]Various combinations of relatedness functions, vector similarity functions, and relatedness thresholds for choosing word sets were used, which resulted in a total of 350 experiments. The purpose of this thoroughness was to discover the most promising combinations of functions for future experiments. It was *not* an effort simply discover a combination that worked. In fact, most combinations performed reasonably well, compared to chance. If all the results of all the experiments were due to chance, the probability of all 350 experiments producing 30 or fewer correct sense-assignments is 0.96. In fact, only 145 of the 350 experiments produced 30 or fewer correct sense-assignments. Thus, successes cannot be attributed to simply capitalizing on chance.

Table 6. Rates of correct sense selection for the experiments with the best performance.

Relatedness Threshold	*Relatedness Function*	*Vector Similarity Function*	*Assignment to Correct Sense*	*Assignment to Correct Group*
none	none (*coc*)	*NDP*	23%	52%
0.1	*cp*	HIT^{x}	45%	79%
0.03	*dcp*	$HIT^{\rightarrow}(R^{C}, R^{S})$	15%	97%

At its best, this method was able to correctly tag 45% of the test sentences. This is a reasonably good performance, given that correctly identifying the exact sense of *bank* in these sentences proved very difficult for the authors. Remember that this method completely ignores syntactic information, including morphology. It is unreasonable to expect any method that does not take syntax into account to reliably distinguish between words that have very similar nominal and verbal forms, such as *bank*. Furthermore, as mentioned above, *bank* has several senses that are very close in meaning.

The technique of expanding contexts and sense-entries to include related words (i.e., words judged to be related according to some relatedness function) proved beneficial. Without expansion, the correct sense assignment was made at best 23% of the time, whereas with expansion the highest rate of correct sense assignment was 45%. The example sentence shown above, which had no words in common with the appropriate sense definition, was generally sense-tagged correctly, demonstrating that the technique can work even under difficult circumstances.

Pathfinder Networks Derived from Co-occurrence Data

One of the problems with co-occurrence data is the sheer quantity of it. There are nearly 2.5 million frequencies of co-occurrence for the words in the LDOCE-controlled vocabulary. Such an enormous amount of data is difficult to use in raw form. However, any reduction must be accomplished without eliminating useful information.

Pathfinder analyses were performed on a relatedness matrix (using the IOU relatedness function) for the 2,177 controlled vocabulary that occur in less than 10% of the text units. The Pathfinder r parameter was always infinity in the analyses reported. Figure 1 shows a fragment of a Pathfinder network including the words within three links of the word *bank*.

The networks that resulted from the Pathfinder analyses were used to select the related sets of words. Related sets were formed by selecting words that were directly linked to a particular word in the network. The number of links connected to each word (the *degree* of the node) varies, depending on the extent to which other words consistently co-occur with it. This means that related-sets also vary in size.

Sense-definition sets were formed for the test word by combining the related-word sets for the words in each sense definition. The union of the related-word sets was used in the experiments described here, although weighting words differentially may be valuable for future investigations of this technique.

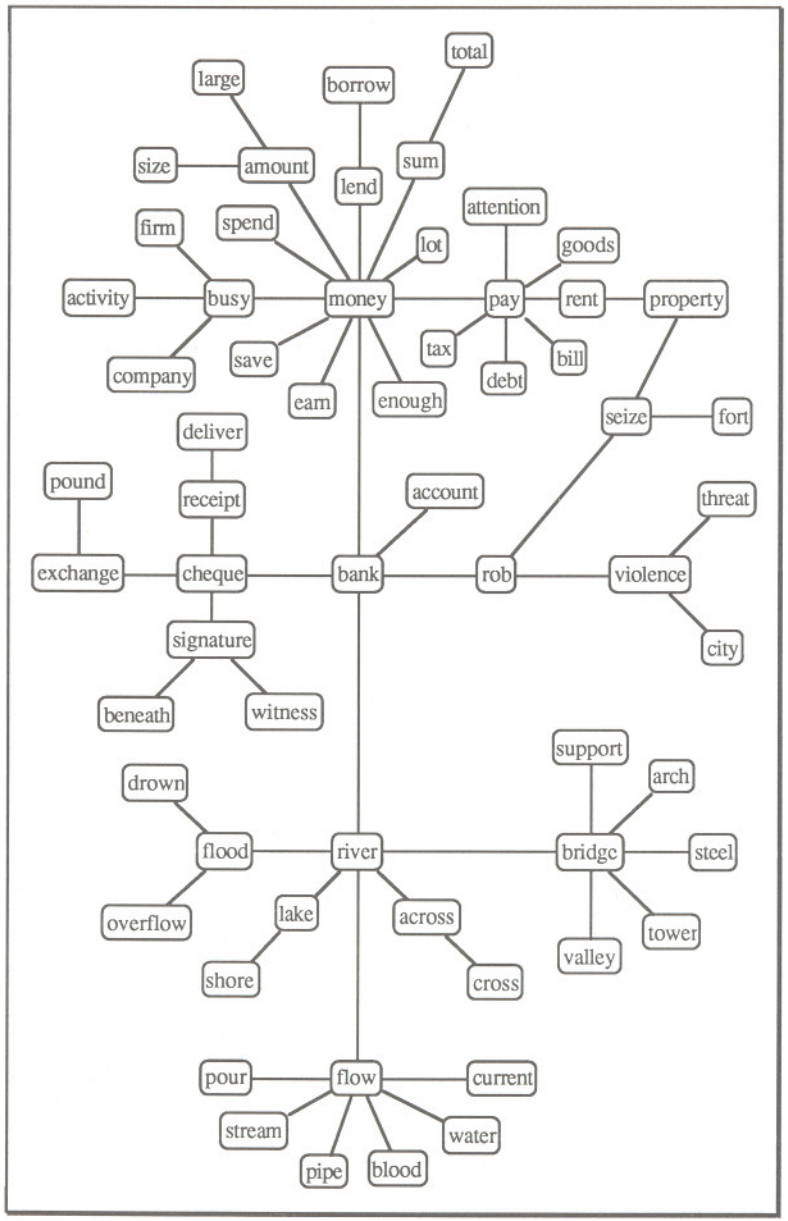

Figure 1. The fragment of the PFNET($r = \infty$, $q = 5$) showing the 61 words within 3 links of *bank*.

As with the sense-definition sets, context sets were formed by combining the related-word sets for all the words in the context sentence. However, our approach here has been to progressively expand the size of the context set by increasing the number of links, or network distance, used in determining relatedness. This is analogous to passing markers in the network starting with the words in the context sentence and continuing in steps passing markers over the links at each step. Finally, a measure of match was computed for each sense definition at each distance from the context set. Although several measures have been considered, the results of using the ratio of the number of words in both context and sense-definition sets to the number of words in either set are reported here.

We used three Pathfinder networks: PFNET(∞,2) with 16,955 links (Q2), PFNET(∞,5) with 3,136 links (Q5), and PFNET(∞,32) with 2,204 links (Q32). With each network, we attempted to identify the correct sense of the word *bank* in the 197 example sentences from LDOCE. In these tests, the sense-definition sets contained only the words in the definitions themselves (no related words, no weights). The context set was progressively expanded by adding the words directly connected to the words in the context set to obtain the Step 1 words, then the words connected to the Step 1 words were added in Step 2, and so forth. At each step, the COM evaluation function was used to compute a strength for each sense definition, and the sense definition with the greatest strength was taken to be the appropriate sense of the word *bank* for the particular context sentence.

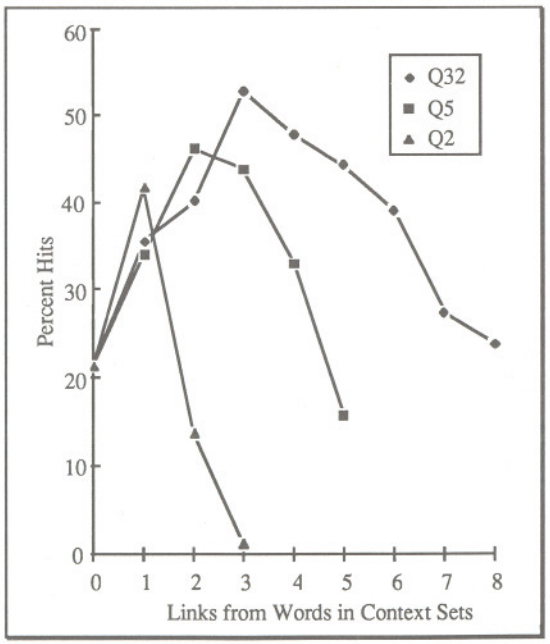

Figure 2. Percent hits using three Pathfinder networks to sense-tag 197 *bank* sentences.

The results of these sense-selection tests are shown in Figure 2. In terms of absolute performance, the network with the fewest links (PFNET(∞,32)) performed best, allowing

bank to be correctly sense-tagged in 104 of the 197 example sentences (53%). Maximum performance occurred when the context set had been expanded to include items three links away (average context-set size = 102). Performance with the PFNET(∞,5) was next best (91 hits at Step 2; average context-set size = 81), and the PFNET(∞,2) was worst (82 hits at Step 1; average context-set size = 91). All of the networks improved on Step 0 performance which is similar to a keyword search using only the words in the context sentence.

The performance of the PFNET(∞,32) network is particularly surprising since it has the fewest links. Apparently, limiting the links to the relatively strongest relations available yields some advantage. It is possible that these links are more immune to the effects of incidental co-occurrences of words in LDOCE that have little to do with the inherent meanings of the words.

Although these results are promising, they may still be of limited practical value. However, the task of choosing the correct sense from a large set of highly similar senses (there are 13 senses for *bank*) may be too stringent a test. Therefore, we also examined performance with the PFNET(∞,32) when only the six sense groups for *bank* were considered. The hit-rate improved to 85% (167 out of 197), a far more usable result. At present, it appears that Pathfinder is capable of capturing the important relationships in the co-occurrence data without losing much of value, at least for our application.

Conclusions

The co-occurrences of words in the LDOCE-controlled vocabulary in the definitions in LDOCE appear to provide some useful information about the meanings of those words. Co-occurrence frequency correlates significantly with human judgments of relatedness, and the relatedness functions on co-occurrences yield even higher correlations. When the relatedness functions are used to derive Pathfinder networks on the LDOCE primitives, these networks serve to represent aspects of the intensional meaning of words. More specifically, one might say that the intensional meaning of a word is represented by the collection of words that are nearby in the network. The experiments on lexical sense selection suggest that the co-occurrence data and the networks derived from those data do capture some aspects of the meanings of words.

Lexical sense selection using co-occurrence data is promising but far from perfect. The Pathfinder networks dramatically reduce the amount of information that must be stored in order to do lexical sense selection. It would be useful to combine information from our method of sense selection with other methods for extracting information from text. Sampson (1986) presents a statistical technique for assigning part-of-speech labels, for example, which would be an excellent candidate for such combination.

At the beginning of this chapter we argued that LDOCE is a relatively good source of co-occurrence data, but that it isn't perfect. To reiterate, LDOCE contains short textual units, each of which is relatively focused. At the same time, the distribution of topics in LDOCE is relatively broad. Importantly, LDOCE is based on a controlled vocabulary, which makes co-occurrence data manageable. However, only a limited number of the senses of the words in the controlled vocabulary are used to define other words in the dictionary, and co-occurrence data cannot reflect relationships involving senses that aren't used. Another potential shortcoming is that LDOCE contains relatively few examples of "definition by synonymy" as compared to other dictionaries. One technique for improving our co-occurrence estimates would be to obtain additional co-occurrence data from other dictionaries and thesauri.