26 From Words to Understanding

Jussi Karlgren and Magnus Sahlgren

As was discussed in section 22, language is central to a correct understanding of the mind. Compositional analytic models perform well in the domain and subject area they are developed for, but any extension is difficult and the models have incomplete psychological veracity. Here we explore how to compute representations of meaning based on a lower level of abstraction and how to use the models for tasks that require some form of language understanding.

26.1 The Meaning of 'Meaning'

The use of vector-based models of information for the purpose of representing word meanings is an area of research that has gained considerable attention over the last decade. A number of different techniques have been suggested that demonstrate the viability of representing word meanings as semantic vectors, computed from the co-occurrence statistics of words in large text data (e.g. Deerwester et al., 1990; Schütze, 1992; Lund and Burgess, 1996). Unfortunately, the philosophical rationale for this practice has remained tacit, which is remarkable since the vector-based models purport to uncover and represent word meanings. What, one might ask, are those meanings that the words of our language apparently have? Are they perhaps some form of mental concepts that exist in the minds of language users, or are they merely the objects named by the words, and if so, how do we represent something like it in a computer system? It seems that we need to know *what* it is we want to represent before we can start thinking about *how* to represent it in a computer system. Succinctly, it seems as if the recourse to semantics demands an explanation of the meaning of 'meaning'.

Ludwig Wittgenstein suggests in *Philosophical Investigations* (1953) that we should view meaning as something founded in linguistic praxis, and that the meaning of a word is determined by the rules of its use within, as he puts it, a specific *language-game*. This suggestion led to the famous dictum "meaning is use," which is sometimes referred to as a Wittgensteinian theory of meaning. The idea is that to understand the meaning of a word, one has to consider its use in the context of ordinary and concrete language behavior. To know the meaning of a word is simply to be able to use the word in the correct way in a specific language-game or linguistic praxis. This line of reasoning thus allows us to define semantic knowledge as that which we *make use of* when successfully carrying out linguistic tasks. According to this way of thinking, meaning is the vehicle by which language travels.

Thus it is language itself and not the concept of meaning that is primary to Wittgensteinian semanticist. The question about the meaning of 'meaning' must therefore be answered "from within" a theory of language, since words do not (and in a stronger sense *cannot*) have meaning outside language. That is, it does not make any sense to ask what the meaning of a word is in isolation from its use in language, since it is only by virtue of this use that the word has meaning. The lack of rigid designations regarding the concept of meaning facilitates our understanding of language as a dynamic phenomenon. What we need in order to understand the nature of meaning is not so much a rigid definition of the concept of meaning, but rather a profound understanding of the inherent structures of natural language. In short, what we need is a structuralistic account of language.

Using such a relatively agnostic theory of meaning, we will in what follows attempt to exemplify its utility for information-access tasks, arguably the most important and applied of language-technology tasks, and one that relies crucially on some form of textual understanding.

26.2 A Case in Point: Information Access

Text is the primary repository and transmitter of human knowledge. Many other types of knowledge representation have been proposed and used for specific purposes, but for most purposes text has proven efficient, flexible, and compact for generation, storage, and access. But while accessing information in text is simple and unproblematic for a human reader, finding the right text to access may be difficult. Computer systems can be of help here, but to do this, systems must have some form of understanding of text content.

26.2.1 System View of Documents

Information-access systems view documents as carriers of topical information and hold words and terms as reasonable indicators of topic. The techniques used for analysis and organization of document collections are focused primarily on word and term occurrence statistics. Documents and information needs alike are analyzed in terms of words.

Although this simple approach has its obviously effective characteristics, it also has some drawbacks. The results provided by information-access systems of today are unimpressive: by the standard metrics defined and practiced in the field, nothing like optimal performance is delivered by any system. To some extent, this is a problem that has to do with the indeterminacy of the evaluation metrics themselves: Relevance is an ill-defined characteristic of documents. But to a great extent systems do not deliver what should and could be expected of them be-

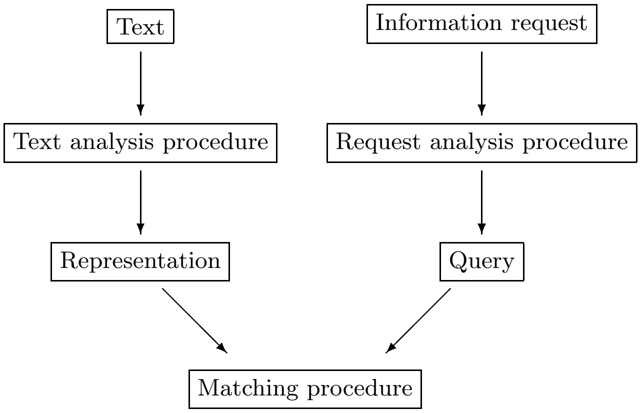

Figure 26.1. The standard model of information retrieval.

cause they model topic and text unsatisfactorily. Although the model is simple and designed not to rely on brittle theory, it does not reflect the underlying structure of textual information transmission sufficiently for purposes of designing useful systems for information access.

26.2.2 The Standard Model for Information Retrieval

The standard model for information retrieval and the basis for most information-system design is roughly as shown in figure 26.1. There is some body of texts; information requests are put to some system that handles this body of texts; the texts are analyzed by some procedure to yield a nontextual representation of them; the information requests are likewise analyzed by an identical or similar procedure to yield a query. The two representations are then matched. The texts with the best matches are presented as potential information sources to fulfill the request.

The point of the analysis is to facilitate matching by (1) reducing the amount of information, to make the representations manageable—it must somehow counter the variability of language and the freedom it affords language users—and (2) resolving the vagueness and indeterminacy inherent in language. The resulting representations are assumed in some way to be alinguistic and amenable to pure formal manipulation.

This quite intuitive and in many ways appealing model hides the complexity of human language use from the matching procedure, which can then be addressed using formal methods. This is not entirely to the benefit of the enterprise. The very same mechanisms that make the matching complicated—the vagueness and indeterminacy of human language—are what make human language work well as a communicative tool. Awareness of this is typically abstracted out of the search process. The major difference between using an automated information-retrieval system and consulting with a human information analyst is that the latter normally does not require the request to be transformed into some invariant and unambiguous representation; neither does the human analyst require the documents themselves to be analyzed into such a representation. A human analyst not only copes with but utilizes the flexibility of information in human language: It is not an obstacle but an asset. For nontrivial retrieval as performed by humans, concepts glide into each other painlessly and with no damage done to the knowledge representation they utilize, and documents that have previously been thought to be of some specific type or topic can be retrieved for perfectly new and unexpected purposes.

26.3 Words as Content Indicators

The basic drawback of the systems of today is their impoverished picture of text content. They treat texts as containers for words, and words as neat and useful indicators of content. But there is no exact matching from words to concepts. Words are vague, both polysemous and incomplete: every word means several things and every thing can be expressed with any one of several words. Words suffer from the combined drawback of being ambiguous and nonexclusive as indicators of content.

Any representation of content should transcend this inherent ambiguity of words—and this is how our kind of model is intended to improve matters.

26.3.1 The Distributional Hypothesis

Viewed from a structuralistic perspective, natural language may be characterized as a sequence of semantically arbitrary symbols (i.e. words). The symbols are semantically arbitrary since it is not their physical properties that determine their meaning. Rather, it is the relations between the symbols of the sequence that define them. This is to say that the meanings of the symbols in the sequence do not derive from any inherent semantic properties of the symbols.

This relational aspect of meaning can be seen if we consider a polysemous word, for example ‘fly’, which has a different meaning in the context of A, for example in texts referring to aerial activities, than in the context of B, for example in texts about insects. The reason the word has different meanings in these different contexts is not that it has a different inherent semantic property in the context of A than in the context of B, for that would mean that a word could change its inherent semantic properties at any time, making linguistic communication, understanding, and hence language itself virtually impossible. The reason for the different meanings of the word is rather that the context of A is different from that of B.

This characterization of natural language as a linear sequence of semantically arbitrary elements allows us to formulate a theory of meaning known as the *distributional hypothesis*. The theory originates from the work of Zellig Harris, who, in his book *Mathematical Structures of Language* (1968, p. 12), states that "the meaning of entities, and the meaning of grammatical relations among them, is related to the restriction on combinations of these entities relative to other entities."

These combinatorial restrictions can be viewed as semantic constraints that govern the distribution of entities in language. That is, if two separate entities occur in combination with the same set of other entities C, the distribution of the two separate entities are governed by the same semantic constraints that are manifested in the distributional pattern that consists in the co-occurrence with C. It is by virtue of this distributional similarity that the entities have similar meaning. This is to say that similarity in distribution implies similar values of semantic information.

The merit of this hypothesis in relation to the Wittgensteinian account of meaning is that the distributional patterns of words can be thought of as manifesting language use. That is, the use of a word is manifested by its distribution in language, which in turn is defined by the contexts in which the word occurs. This means that the context could be utilized as a measure of distribution, by which the use (and thus also the meaning) of a word could be determined. Thus, if the *meaning* of a word is determined by its *use* in language, and its use is manifested by its *distribution*, the distributional patterns as defined by the *contexts* of a word can be seen as viable tools for determining the meaning of that word.

The idea is that words are semantically similar to the extent that they share contexts. If two words w_1 and w_2, say 'beer' and 'wine', frequently occur in the same context C, say after 'drink', the hypothesis states that w_1 and w_2 are semantically related, or, stated more strongly, that they are semantically similar. The semantic similarity (or relatedness) of 'beer' and 'wine' is thus due to the similarity of usage of these words. This means that the categorization of both words as, for example, referring to alcoholic beverages is possible only because we use them in such a way; for example, after the word 'drink' and in the vicinity of the word 'drunk'. The categorization is not a *cause* of usage but a *consequence*.

In an attempt to formalize these ideas, we could say that the meaning

M of a word w is determined by its distribution D in text T. D over T can be defined as the union of the contexts C in which w occurs. Thus, if $D(w,T)$ (the distribution of w in T) determines $M(w)$ (the meaning of w), and $D(w,T)$ equals $\sum\{C \mid w \in C\}$ (the contexts in which w occurs), this can be seen as a representation of $M(w)$. This representational scheme thus justifies the claim that word meanings *can* be uncovered and represented in computer systems. It also justifies the claim that this can be done without forcing us to commit to any particular ontology about *what* these meanings are. Rather, it is because we do not demand any rigid definition of the concept of meaning that this representational scheme becomes feasible.

26.4 Latent Semantic Analysis

This representational scheme is the communal rationale for vector-based semantic analysis. The assumption that "words with similar meanings will occur with similar neighbors if enough text material is available" (Schütze and Pedersen, 1997) is central to all the various approaches. What separates them is *how* they implement the idea. The pioneering technique in this area of research, latent semantic analysis (LSA) (Landauer and Dumais, 1997), collects the text data in a words-by-documents co-occurrence matrix where each cell indicates the frequency of a given word in a given text sample of approximately 150 words. The words-by-documents matrix is normalized by using logarithms of word frequencies and entropies of words across all documents. The normalized matrix is then transformed with singular-value decomposition into a much smaller matrix. This dimension reduction appears also to accomplish inductive effects, reminiscent of human psychology, by capturing latent semantic structures in the text data. Words (or, more to the point, *concepts*) are thus represented in the reduced matrix by semantic vectors of dimensionality n (300 proving to be optimal in Landauer and Dumais's 1997 experiments).

A similar approach is taken by Schütze and Pedersen (1997), who represent the text data as a words-by-words co-occurrence matrix where each cell records the number of times that the word pair occurs in a window spanning 40 words. However, such a matrix with $v^2/2$ distinct entries, where v is the size of the vocabulary, becomes computationally intractable for large vocabularies, so they first approximate the matrix using class-based generalization in two steps and then transform it with singular-value decomposition, so that words are represented in the final reduced matrix by dense semantic vectors of dimensionality n ($n = 20$ in Schütze and Pedersen, 1997). The motivation for reducing the dimensionality of the approximated matrix with singular-value decomposition is, as in LSA, that it improves generalization and makes the representations more compact.

This kind of representational scheme where words are represented as semantic vectors that are calculated from the co-occurrence statistics of words in large text data has proven to be both computationally advantageous and cognitively justified. The drawback of singular-value decomposition is that it places heavy demands on computing time and memory, which suggests that an alternative way of achieving the inductive effects of dimension reduction might be worth considering. A number of techniques for doing so have been proposed under such names as random mapping (Kaski, 1998), random projections (Papadimitriou et al., 1998), and random indexing (Kanerva et al., 2000), and they have the same underlying mathematics.

26.5 Random Indexing

In previously reported experiments with random indexing (Kanerva et al., 2000), documents of approximately 150 words each are represented as high-dimensional random index vectors (dimensionality $> 1{,}000$) that are accumulated into a words-by-documents matrix by adding a document's index vector to the row for a given word every time the word appears in that document. The method is comparable to LSA, except that the resulting matrix is significantly smaller than the words-by-documents matrix of LSA, since the dimensionality of the index vectors is smaller than the number of documents. By comparison, assuming a vocabulary of 60,000 words in 30,000 documents, LSA would represent the data in a $60{,}000 \times 30{,}000$ words-by-documents matrix, whereas the matrix in random indexing would be $60{,}000 \times 1{,}800$ when 1,800-dimensional index vectors are used. This seems to accomplish the same inductive effects as those attained by applying singular-value decomposition to the much larger matrix, but without the heavy computational load that singular-value decomposition requires.

In the present experiment, the high-dimensional random vectors of random indexing have been used to index *words* and to accumulate a words-by-contexts matrix by means of *narrow* context windows consisting of only a few adjacent words on each side of the focus word. As an example, imagine that the number of adjacent words in the context window is set to two. This would imply a window size of five space-separated linguistic units, i.e. the focus word and the two words preceding and succeeding it. Thus the context for the word 'is' in the sentence 'This parrot is no more' is 'This parrot' and 'no more', as denoted by

[(This parrot) is (no more)].

Calculating semantic vectors using random indexing of words in narrow context windows is done in two steps. First, an n-dimensional sparse random vector called a *random label* is assigned to each word type in the text data. These labels have a small number k of randomly distributed -1s and $+1$s, with the rest set to 0. The present experiment utilized 1,800-dimensional labels with $k = 8.7$ on average, with a standard deviation of ± 2.9. Thus a label might have, for example, four -1s and six $+1$s. Next, every time a given word—the focus word f_n—occurs in the text data, the labels for the words in its context window are added to its *context vector*. For example, assuming a $2 + 2$ sized context window as represented by:

$$[(w_{n-2}w_{n-1})f_n(w_{n+1}w_{n+2})]$$

the context vector of f_n would be updated with:

$$L(w_{n-2}) + L(w_{n-1}) + L(w_{n+1}) + L(w_{n+2})$$

where $L(x)$ is the label of x. This summation has also been weighted to reflect the distance of the words to the focus word. The weights were distributed so that the words immediately preceding and succeeding the focus word get more significance in the computation of the context vectors. For the four different window sizes used in these experiments, the window slots were given weights as follows:

1 + 1:	[(1) 0 (1)]
2 + 2:	[(0.5, 1) 0 (1, 0.5)]
3 + 3:	[(0.25, 0.5, 1) 0 (1, 0.5, 0.25)]
4 + 4:	[(0.1, 0.1, 0.1, 1) 0 (1, 0.1, 0.1, 0.1)].

The rationale for these operations is that a high-dimensional context vector, by effectively being the sum of a word's local contexts, represents the word's relative meaning. This means that we will be able to model word content with some confidence. But texts are more than words and their content.

26.6 What Is Text, from the Perspective of Linguistics?

The model described above covers much of what we want in terms of human linguistic behavior. But what we know about language and text certainly motivates a more sophisticated model than set theory on the level of word occurrences. Linguists treat linguistic expressions as being composed of words that form clauses that in turn form text or discourse. Words have predictable situation-, speaker-, and topic-*independent* structure that is described formally. Clauses have largely predictable situation-, speaker-, and topic-*independent* structure that is described formally. This is how far formal linguistic analysis takes us. Attempts at formal analysis at the next level—of text and topic structure—have been only partially successful. Texts have largely unpredictable situation-, speaker-, and topic-*dependent* structure, which cannot be handled adequately with the theoretical apparatus available to us today. Clause structure is connected only indirectly to topicality: mostly it accounts for the local organization of the clause. However, the invariant and predictable nature of clause structure certainly encourages further attempts at building theories that relate meaning to clause structure, and it would be foolish to build a text model that can take no account of recent advances in formal analysis of text structure.

26.6.1 Beyond Word Co-occurrence—Implementing Linguistics

The only property of text that is being utilized in the creation of the context vectors is the distributional patterns of linguistic entities. This comprises, however, only a small fraction of the structural complexity of large texts. There are other inherent structural relations in natural language that might be significant for uncovering semantic information. The distributional hypothesis does tell a story about the foundation of meaning, but it might not tell the *whole* story. If the overall goal of the research is to understand how meaning resides in language, and how to implement linguistic knowledge about meaning in computers, it seems unmotivated not to take these more complex linguistic features into account. Therefore, we have evaluated the method using different degrees of linguistic preprocessing of the training data, such as morphological analysis and part-of-speech tagging, with the intention of investigating whether the utilization of more sophisticated linguistic information in some way concretizes the semantic information captured in the context vectors. To ensure state-of-the art performance in linguistic analysis, we used the functional dependency grammar of English—the FDG parser—developed by Conexor to analyze the text and its words (Järvinen and Tapanainen, 1997).

26.7 The TOEFL-Test

To repeat the fundamental hypothesis of this investigation, the *raison d'étre* of the high-dimensional context vectors described in above sections is that they represent the relative meaning of words, and that they therefore can be used to calculate semantic similarities between words.

We will verify this hypothesis by letting the system perform a synonym test. One such test is TOEFL (test of English as a foreign language), which is a standardized test employed, for example, by American universities to evaluate foreign applicants' knowledge of the English language. In the synonym-finding part of the test, the test taker is asked to find the synonyms to certain words. For each given word, a choice from four alternatives is provided, where one is the intended synonym and is supposed to be indicated by the person taking the test. In the present experiment, 80 test items of this type were used.

When performing the synonym test, the system simply calculates the distances (the cosine of the angle between context vectors) between the target word and the four alternatives and gives the highest-ranking alternative as its answer (i.e. the one that correlates most closely with the target word).

26.8 Experimental Set-Up

The text data used as learning material for the system was a ten-million-word corpus of unmarked English with a vocabulary of 94,000 words. In the first stage of preprocessing the number of unique word types was reduced based on frequency, by weighting the least and the most frequent words with 0 when they appeared in the context window. A frequency range of 3–14,000 was used and resulted in a vocabulary of 51,000 words. Next, a rather crude method for morphological analysis was implemented by truncating the words. The idea was to approximate word stems by simply chopping off the words at a certain predefined number of letters. In the present experiment, truncation lengths of 6, 8, 10, and 12 were used. As a comparison to the crude truncation approach to morphology, the the Conexor FDG parser was used to analyze the text initially and extract the base form of each word.

The Conexor FDG parser was also used to supply the analyzed text (the text consisting of proper word stems) with part-of-speech information in an attempt to deal with the ever-present ambiguity of languages, the problem being that the same "word" can have several meanings, and many of these orthographically identical but semantically dissimilar words belong to different parts of speech. For example, *roll* can be used as a verb or as a noun. Providing part-of-speech information by simply adding the part-of-speech tag to the beginning of each word would enable the system to detect this kind of ambiguity and to discriminate between these words. For example, the verb *roll* would become *vroll*, whereas the noun *roll* would become *nroll*.

26.9 Results and Analysis

The results of the TOEFL-test are summarized in table 26.1. The numbers in the cells are averages over five runs. The standard deviation for these results is ±1.5. All results are given in percent of correct answers to the TOEFL-test. The numbers in boldface are the results from ex-

Table 26.1

Average Results (±1.5) in Percent of Correct Answers to the TOEFL-test Tr. means truncation length, **WS** means 'word stems', and **PoS+WS** means 'part-of-speech tagged word stems'.

Linguistic	Context window				Average
analysis	$1+1$	$2+2$	$3+3$	$4+4$	(±0.73)
None	64.5	67.0	65.3	65.5	65.6
Tr. 6	55.0	57.5	57.3	55.3	56.3
Tr. 8	61.5	64.3	62.0	63.3	62.8
Tr. 10	66.0	68.5	66.3	66.3	66.8
Tr. 12	64.8	65.3	63.8	64.8	64.6
WS	**63.5**	**70.8**	**72.0**	**66.0**	**68.1**
PoS+WS	**66.0**	**64.5**	**65.0**	**65.5**	**65.3**
Average (±0.56)	63.0	65.4	64.5	63.8	

periments with linguistically analyzed text. By comparison, tests with LSA on the same text data, using the LSIBIN program from Telecordia Technologies, produced top scores at 600 factors of 58.75% using the unnormalized words-by-documents matrix, and 65% using a normalized one. The average result reported by Landauer and Dumais (1997) with LSA (using normalization and different text data) is 64.4%, while foreign (non-English-speaking) applicants to U.S. colleges average 64.5%.

These results indicate that high-dimensional random labeling of words in narrow context windows captures similarity relations between words just as effectively as singular-value decomposition of the words-by-documents matrix does (e.g. LSA), as measured by a standardized synonym test. Without using linguistic information, the system averages 65.6% over the four different window sizes used in these experiments. However, already when utilizing a rather naive kind of morphological analysis in the form of carefully applied truncation (using a truncation length of 10 characters), the system's average result increases to 66.8% correct, although it seems imperative not to truncate too early, since this gravely affects the results. Shortening the truncation length to eight characters decreases the result to 62.8%, and shortening it to six renders a meager 56.3%. Extending the truncation length to twelve characters also decreases the result, with a 64.6% average.

The best results were produced by proper stemming of words, which yields 68.1% correct on the average. This indicates that since the inclusion of morphology in the form of proper word-stem analysis or carefully applied truncation yields the best overall results, taking advantage of other inherent structural relations in text, in addition to the distributional patterns of linguistic entities, really might be significant for uncovering semantic information from text data. However, adding part-of-speech information did not further improve the performance. The average result when adding part-of-speech information to the morphologically analyzed text drops to 65.3%. This could be a result of the increase in the size of the vocabulary, which is the consequence of supplying part-of-speech information for each word.

Turning now to the different window sizes, the table shows that a minimal context window with just one word on each side of the focus word yields the worst average result. This might not be surprising, since it seems reasonable to assume *a priori* that a minimal context window will not provide enough contextual information for making the comparison of distributional similarity reliable. This assumption is not categorically supported by the results, however, since for the part-of-speech-tagged word stems, a 1 + 1 sized context window actually produces the best average result. The results peak in the range of two to three words on each side of the focus word, with a 2 + 2 sized context window producing the best average result of 65.4%, but with a 3 + 3 sized context window producing the best individual result of 72% using the morphologically analyzed text. A 4 + 4 sized context window is only slightly better (63.8%) than the minimal context window, and our experiments with context windows exceeding four words on each side of the focus word gave much lower scores.

26.10 Some Cognitive Implications

The results from these experiments demonstrate that the technique is capable of achieving comparatively good results on a standardized synonym test. The test is designed to measure word knowledge, which would indicate that any subject capable of performing the test with scores above the level of guessing (which statistically would yield 25% correct) possesses a certain amount of linguistic knowledge about word meanings. Therefore, the test could also be seen as a rudimentary intelligence test. Landauer et al. (1998) point out that "word-word meaning similarities are a good test of knowledge—indeed, vocabulary tests are the best single measure of human intelligence."

The results achieved in these investigations are approximately parallel to the results accomplished by foreign applicants to American universities. The question is, then, if the results, viewed from this perspective, justify the conclusion that the system has acquired and applied linguistic knowledge (about word meanings)? Do the high-dimensional random distributed representations constitute a viable model of semantic knowledge? In short: What are the cognitive implications of the accomplishments of the system?

The keyword in this discussion is *functionality.* The performance of the system could be described as *functionally* equivalent to the linguistic behavior of a human language user in carrying out the specific predefined linguistic task of picking out synonyms of a target word. This means that since the system's internal representations in the form of context vectors have been proven (by the successful execution of the TOEFL-test) to be functional for purposes pertaining to linguistic competence, we may describe the system as having acquired and applied the *computational equivalent* of the linguistic knowledge that humans possess when discriminating between word meanings.

This characterization of the system means that the relevant question is *not* whether the system's internal representations of word meanings actually *mean* anything (i.e. if they somehow correspond to how word meanings are represented in the human mind—assuming this question is meaningful), but rather whether they can be *utilized* for the purpose of modeling observable linguistic behavior. The semantic information that the context vectors carry does not reside in the vector representations alone, but rather in the relations between the vectors. The representation is relative rather than absolute, since it is only in relation to each other that the context vectors *mean* anything. The important point is therefore that the system's internal representations can produce linguistic behavioral patterns that manifest semantic knowledge—and that can be regarded a fragment of the *functionality* of a language user. In other words: It is by virtue of letting the meaning of 'meaning' remain indeterminate that we may consider the implementation of a functional pattern as an epistemic or cognitive achievement.

26.11 Implications for Information Access

More concretely, for immediate attention, if we wish to improve on information-access systems—if we use the standard architecture as delineated in the beginning of this chapter—there are three access points, points where the character of the internal text representation influences the working of the system as a whole, and points where a more adaptive and humanlike processing would make a difference:

1. the intelligent selection of document descriptors for each document;
2. the flexible internal expression of those items; and
3. the negotiable elicitation of information needs from the reader.

And this functionality should not be restricted to a single language.

26.12 Meaning in Text

The reason for using narrow context windows to calculate semantic word vectors as opposed to using whole documents, as in LSA, is the assumption that the semantically most significant context is the immediate vicinity of a word. That is, one would expect the words closest to the focus word to be more important than the words further away in the text for determining the meaning of the focus word. The intuition is that a local context is more reliable for measuring semantic similarity between words than a large context region spanning hundreds of words.

This intuition is expressed, for example, by Lin (1997), who states that "two different words are likely to have similar meanings if they occur in identical local contexts." Schütze and Pedersen (1997) argue that local co-occurrence statistics are both qualitatively and quantitatively more informative than document-based co-occurrence statistics, since the number of co-occurrence events will be higher when using a sliding window to define the co-occurrence region than when using documents, especially if the documents are long. Burgess and Lund (2000) also report on the merits of narrow context windows.

If the assumption is correct that local co-occurrence statistics give a more reliable measure of distributional similarity between words than do document-based co-occurrences, one should be able to discern an increase in performance when using narrow context windows, as opposed to documents, for calculating the semantic word vectors. The results of our experiments seem to favor this assumption. Compared to the performance of techniques based on context regions in excess of a hundred words, such as LSA, narrow context windows perform well in at least one linguistic task (TOEFL) pertaining to lexical semantic knowledge.

This improved performance raises the question of whether there might be a difference in what sort of semantic information can be extracted by considering different amounts of context. A larger context might give better clues to what a particular word is about than to what it means. The idea is that two words that are about similar things will occur in similar context regions (e.g. documents), while two words that have similar meanings will occur with similar context neighbors (i.e. words). This means that larger contexts might be more suited for tasks pertaining to topical information, such as information retrieval, than in tasks directed specifically toward lexical semantic competence. The applicability of LSA to information retrieval is well documented (e.g. Dumais et al., 1988; Deerwester et al., 1990), supporting this assumption.

The possibility of a discrepancy between the kinds of semantic information carried by different context sizes suggests that although a synonym test is a fairly reliable method for measuring one kind of semantic knowledge, other conceivable methods for measuring semantic knowledge might be worth considering. Other evaluation procedures have been reported in the literature, such as comparing vector similarities with reaction times from lexical priming studies (Lund and Burgess, 1996) or using LSA for evaluating the quality of content of student essays on given topics (Landauer et al., 1997).

Meaning, the main object of our study, is most decidedly situation-dependent. While much of meaning appears to achieve consistency across usage situations, most everything *can* be negotiated on the go. Human processing appears to be flexible and oriented toward learning from prototypes rather than learning by definition: Learning new words and adding new meanings or shades of meaning to an already known word do not require a formal retraining process. And, in fact, natural use of human languages does not make use of definitions or semantic delimitations; finding an explicit definition in natural discourse is a symptom of communicative malfunction, not of laudable explicitness.

A text model should model language *use* rather than language in the abstract. We need a better understanding of how meaning is negotiated in human language usage: Fixed representations do not seem practical and do not reflect observed human language usage. We need a more exact study of inexact expression, of the *homeosemy* ('homeo' from Greek *homoios* similar) or near and close synonymy of expressions of human language. This means we need to understand the temporality, saliency, and topicality of terms, relations, and grammatical elements—it means modeling the life cycle of terms in language, the life cycle of referents in discourse, and the connection between the two. These experiments have taken but some first steps in that direction.

References

Aleksander, I. and Morton, H. (1995). *An Introduction to Neural Computing*, second edition. London: International Thomson Computer Press.

Boden, M. B. and Niklasson, L. F. (1995). Features of distributed representation for tree structures: A study of RAAM. In L. F. Niklasson and M. B. Boden (eds.), *Current Trends in Connectionism.* Hillsdale, NJ: Erlbaum.

Burgess, C. and Lund, K. (2000). The dynamics of meaning in memory. In E. Dietrich and A. B. Markman (eds.), *Cognitive dynamics: Conceptual change in humans and machines.* Mahwah, NJ: Lawrence Erlbaum Associates.

Chalmers, D. J. (1990). Syntactic transformations on distributed representations. *Connection Science* 2(1–2):53–62.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the Society* *for Information Science* 41(6):391–407.

Dumais, S. T., Furnas, G. W., Landauer, T. K., and Deerwester, S. (1988). Using Latent Semantic Analysis to improve information retrieval. *Proceedings of CHI'88: Conference on Human Factors in Computing* (pp. 281–285). New York: ACM.

Durrett, R. (1995). *Probability: Theory and Examples.* 2nd ed. Belmont, Calif.: Duxbury Press, Wadsworth Publishing Co.

Eliasmith, C, and Thagard, P. (2001). Integrating structure and meaning: A distributed model of analogical mapping. *Cognitive Science* 25(2):245–286.

Fodor, J. A. and Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition* 28:3–71.

Gayler, R. W. and Wales, R. (1998). Connections, binding, unification, and analogical promiscuity. In K. Holyoak, D. Gentner, and B. Kokinov (eds.), *Advances in Analogy Research: Integration of Theory and Data from the Cognitive, Computational, and Neural Sciences* (Proc. Analogy '98 workshop, Sofia), pp. 181–190. Sofia: New Bulgarian University.

Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive Science* 7(2):155–170.

Gentner, D., Holyoak, K. J., and Kokinov, B. K. (eds.) (2001). *The Analogical Mind: Perspectives from Cognitive Science*. Cambridge, MA: MIT Press.

Gentner, D. and Markman, A. B. (1993). Analogy—watershed or Waterloo? Structural alignment and the development of connectionist models of analogy. In C. L. Giles, S. J. Hanson, and J. D. Gowan (eds.), *Advances in Neural Information Processing Systems 5* (NIPS '92, pp. 855–863). San Mateo, CA: Morgan Kaufmann.

Harris, Z. (1968). *Mathematical Structures of Language*. New York: Interscience Publishers.

Hinton, G. E. (1990). Mapping part-whole hierarchies into connectionist networks. *Artificial Intelligence* 46(1–2):47–75.

Hofstadter, D. R. (1984). *The Copycat Project: An Experiment in Nondeterminism and Creative Analogies.* AI Memo 755, Artificial Intelligence laboratory, Massachusetts Institute of Technology.

Hofstadter, D. R. (1985). *Metamagical Themas: Questions of the Essence of Mind and Pattern.* New York: Basic Books.

Jaeckel, L. A. (1989a). *An alternative design for a Sparse Distributed Memory.* Report RIACS TR-89.28. Research Institute for Advanced Computer Science, NASA Ames Research Center.

Jaeckel, L. A. (1989b). *A class of designs for a Sparse Distributed Memory.* Report RIACS TR-89.30, Research Institute for Advanced Computer Science, NASA Ames Research Center.

Järvinen, T. and Tapanainen, P. (1997). *Functional Dependency Grammar.* Publications of the Department for General Linguistics, University of Helsinki.

Kanerva, P. (1988). *Sparse Distributed Memory.* Cambridge, Mass.: MIT Press.

Kanerva, P. (1992). Associative-memory models of the cerebellum. In I. Aleksander and J. Taylor (eds.), *Artificial Neural Networks, 2* (Proc. ICANN '92, Brighton, UK, pp. 23–34). Amsterdam: Elsevier.

Kanerva, P. (1993). Sparse Distributed Memory and related models. In M. H. Hassoun (ed.), *Associative Neural Memories.* New York: Oxford University Press.

Kanerva, P. (1996). Binary spatter-coding of ordered K-tuples. In C. von der Malsburg, W. von Seelen, J. C. Vorbrüggen, and B. Sendhoff (eds.), *Artificial Neural Networks* (Proc. ICANN'96, Bochum, Germany, pp. 869–873). Berlin: Springer.

Kanerva, P., Kristofersson, J., and Holst, A. (2000). Random Indexing of text samples for Latent Semantic Analysis. In L. R. Gleitman and A. K. Josh (eds.), *Proceedings of the 22nd Annual Conference of the Cognitive Science Society,* p. 1036. Mahwah, New Jersey: Erlbaum.

Kaski, S. (1998). Dimensionality reduction by random mapping: Fast similarity computation for clustering. *Proceedings of the IJCNN'98, International Joint Conference on Neural Networks* (Anchorage, vol. 1, pp. 413–418). Piscataway, NJ: IEEE Press.

Kristoferson, J. (1995). *Best probability of activation and performance comparisons for several designs of Sparse Distributed Memory.* Report SICS R95:09, Swedish Institute of Computer Science.

Kristoferson, J. (1997). *Some results on activation and scaling of Sparse Distributed Memory.* Report SICS R97:04, Swedish Institute of Computer Science.

Landauer, T. K. and Dumais, S. T. (1997). A solution to Plato's problem: The Latent Semantic Analysis theory of acquisition, induction and representation of knowledge. *Psychological Review* 104(2):211–240.

Landauer, T. K., Laham, D., Rehder, B., and Schreiner, M. E. (1997). How well can passage meaning be derived without using word order? A comparison of Latent Semantic Analysis and humans. In M. G. Shafto and P. Langley (eds.), *Proceedings of the 19th annual meeting of the Cognitive Science Society,* pp. 412–417. Mawhwah, NJ: Erlbaum.

Landauer, T. K., Laham, D., and Foltz, P. W. (1998). Learning human-like knowledge by Singular Value Decomposition: A progress report. In M. I. Jordan, M. J. Kearns and S. A. Solla (eds.), *Advances in Neural Information Processing Systems* 10, pp. 45–51. Cambridge, Mass.: MIT Press.

Lin, D. (1997). Using syntactic dependency as local context to resolve word sense ambiguity. *Proceedings of ACL-97*, Madrid, Spain.

Lund, K. and Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments & Computers* 28(2):203–208.

Mitchell, M. (1993). *Analogy-Making as Perception: A Computer Model.* Cambridge, MA: MIT Press.

Papadimitriou, C. H., Raghavan, P., Tamaki, H., and Vempala, S. (1998). Latent Semantic Indexing: A probabilistic analysis. *Proceedings of the 17th*

ACM Symposium on the Principles of Database Systems, pp. 159–168. ACM Press.

Plate, T. A. (1994). *Distributed Representation and Nested Compositional Structure*. Ph.D. Thesis. Graduate Department of Computer Science, University of Toronto.

Pollack, J. P. (1990). Recursive distributed representations. *Artificial Intelligence* 46(1–2):77–105.

Rachkovskij, D.A. and Kussul, E.M. (2001). Binding and normalization of binary sparse distributed representations by Context-Dependent Thinning. *Neural Computation* 13(2):411–452.

Rumelhart, D. E, and McClelland, J. L. (1986). On learning the past tenses of English verbs. In J. L. McClelland and D. E. Rumelhart (eds.), *Parallel Distributed Processing 2: Applications* (pp. 216–271). Cambridge, Mass.: MIT Press.

Schütze, H. (1992). Dimensions of meaning. *Proceedings of Supercomputing* (Minneapolis. pp. 787–796). Los Alamitos, CA: IEEE Computer Society Press.

Schütze, H. and Pedersen, J. O. (1997). A cooccurrence-based thesaurus and two applications to information retrieval. *Information Processing and Management* 33(3):307–318.

Sjödin, G. (1995). *Improving the capacity of SDM.* Report R95:12, Swedish Institute of Computer Science.

Sjödin, G. (1996). Getting more information out of SDM. In C. von der Malsburg, W. von Seelen, J. C. Vorbrüggen, and B. Sendhoff (eds.), *Artificial Neural Networks—Proc. ICANN '96,* 477–482. Berlin: Springer.

Sjödin, G. (1997). The Sparchunk code: A method to build higher-level structures in a sparsely encoded SDM. *Proc. 1998 IEEE International Joint Conference on Neural Networks* (IJCNN/WCCI, Anchorage, Alaska, May 1998, vol. 2, pp. 1410–1415). Pincataway, NJ: IEEE Press. Reprinted in this volume.

Sjödin, G., Karlsson, R., and Kristoferson, J. (1997). Algorithms for efficient SDM. *Proc. 1997 Real World Computing Symposium* (RWC'97, Tokyo, January 1997), 215–222. Report RWC TR-96001, Real World Computing Partnership, Tsukuba Research Center, Japan.

Smolensky, P. (1990). Tensor product variable binding and the representation of symbolic structures in connectionist systems. *Artificial Intelligence* 46(1–2):159–216.

Wittgenstein, L. (1953). *Philosophical Investigations.* Translated by G. E. M. Anscombe. Oxford: Blackwell. Swedish translation *Filosofiska undersökningar* by A. Wedberg (1992). Stockholm: Thales.