Chapter 1

Disambiguation and Connectionism

This chapter looks at word sense disambiguation from the perspective of the field of connectionism [27, 22]. We will bring two central concerns in connectionism to bear on designing a disambiguation algorithm: psychological plausibility and large-scale applicability. Several studies suggest that the concept of activation is crucial for explaining human word sense disambiguation (Kawamoto 1988; Tanenhaus, Leiman and Seidenberg 1979; Swinney 1979). Activation is often modelled using vectors in a multi-dimensional real-valued vector space. We will propose that both senses and contexts of use of the ambiguous word be represented as directions in space, and that a context of use is assigned to the sense whose direction it coincides with. The important role of activation in explaing human word sense disambiguation motivates us to adopt such a vector space representation for contexts and senses.

Another aspect of psychological plausibility considered important in connectionism is acquisition. In general, connectionist modelling tries to do without innate constraints or a large body of a priori knowledge unless there is good empirical evidence for hereditary or environmental structure that guides learning. The disambiguation algorithm proposed here acquires knowledge by grouping semantically similar contexts into *context groups.* The assessment of semantic similarity is one of the basic building blocks of human cognition. No psychologically implausible assumptions about strong innate or environmental structure are therefore necessary for our algorithm.

A second major concern of the connectionist enterprise has been to tackle problems of a realistic size. Neural network architectures often fit mathematical models that have been the bread and butter of statisticians for decades. For example, a neural network with no hidden layers, one output node and a sigmoid activiation function corresponds to the well-known logistic model. However, connectionists pay particular attention to problems that occur when models are trained on large data sets. On-line learning and gradient descent algorithms can accommodate data sets that are difficult to handle with traditional statistical fitting methods. In the connectionist tradition of research, we would like to make use of very large data sets, large text corpora in the case of disambiguation. The algorithm proposed here can perform the clustering of very large sets of contexts into groups and therefore exploit the information available in large text corpora that are increasingly available to the public. As proof of the large-scale applicability of the algorithm we will present an application that retrieves information from a large text corpus using disambiguation.

The chapter is organized as follows. Section 1.1 surveys some of the connectionist literature relevant to disambiguation, motivates the use of vector spaces for representing words, contexts and senses, and shows how the acquisition of senses (a problem widely neglected in other work) can be modelled as clustering. Section 1.2 introduces context-group discrimination, a disambiguation algorithm that satisfies the desiderata of representing activation levels, offering a model of acquisition and large-scale applicability. Section 1.3 describes an application to information retrieval which demonstrates that the algorithm can be used for problems of a large size. Section 1.4 discusses context-group discrimination and connectionism in the context of this volume.

1.1 Disambiguation in Connectionism

The importance of the concept of activation in disambiguation was demonstrated in experiments by Tanenhaus et al. [36] and Swinney [35]. They showed that word senses are not selected instantaneously in sentence comprehension. Rather, both senses of words with a two-way ambiguity are activated initially. The sense that receives contextual support remains active, whereas the contextually inappropriate sense is gradually deactivated over the course of several hundred milliseconds. Activation of a sense is determined by testing the extent to which subjects are primed for it. For example, if the word "flower" is displayed on a screen, subjects read it more quickly after having heard sentence 1a than after sentence 1b.

(1) a. They held the rose.

b. They all rose.

The two studies found that reaction times to words related to *both* senses of the ambiguous words were shortened, suggesting that all senses of an ambiguous word are activated initially.

Activation-based models have subsequently been exploited by several connectionist researchers working on disambiguation and language comprehension. McClelland et al. [21] represent sentence meanings as vectors of activations, in this case activations of nodes in a neural network. Given an input sentence, a neural network computes the activation vector that represents the sentence meaning. A second neural network makes inferences from this activation vector. For example, it infers that "busdriver" is the agent in 2a and the patient in 2b.

(2) a. The busdriver was given the rose by the teacher.

b. The busdriver gave the rose to the teacher.

This model can accommodate the gradual deactivation of senses since the sentence activation vector reflects gradual changes of the most likely sentence meaning. Consider as an example the word "bat". When processing the beginning of sentence 3 ("I saw the bat ..."), both senses of "bat" are activated. Further processing ("... flying ...") deactivates the implement sense since the implement is less likely to fly than the animal. After processing “cave” the network settles on the animal sense because caves are not a natural context for the implement.

(3) I saw the bat flying through the cave.

While touching on disambiguation, the model by McClelland et al. [21] deals with sentence comprehension in general. Two early connectionist models by Cottrell [3] and Kawamoto [15] focus explicitly on disambiguation. Both models stress the integration of different forms of information: spelling, pronunciation, conceptual meaning and case information (the latter being information about whether the role of a noun phrase in the sentence is as an actor, patient, instrument, etc.). The information is entered by activating the relevant input nodes of the connectionist network. The activation is then passed on to the other nodes in the network until a stable activation pattern has been reached. In Cottrell’s case, the network disambiguates by choosing the concept node corresponding to the correct sense. This will be the concept node with the highest activation. In Kawamoto’s case, the disambiguation result can be read off of the final distributed pattern of activation. Each concept is represented by a distinct pattern. The patterns of the concepts are compared with the network’s final activation pattern, and the closest concept pattern is chosen as the correct meaning.

In these two models, vector representations (or patterns of activation) are used to integrate multiple sources of activation about word senses. Here, the vector space functions as a common representational medium that allows smooth integration of the available information sources.

Catherine Harris is another connectionist researcher who has worked on representing meaning as patterns of activations in a connectionist network [10]. She focuses on subregularities, i.e., regularities that are intermediate between exceptions and exceptionless rules. An example are English past tense forms like “cut - cut - cut” and “hit - hit - hit” which are memorized as a cluster of verbs with similar morphology, even though not all one-syllable verbs with a syllable-final "t" follow this pattern (e.g., "sit - sat - sat"). In Harris' study, the continuous nature of real-valued representations make the representation of subregularities possible. This is exploited for the disambiguation of the preposition "over", for example the sense "extending over the boundaries of X" ("he stood over the manhole") vs. "on the other side of" ("he lives over the hill").

The above results (gradual deactivation, information integration, and linguistic subregularity) motivate models of disambiguation in a vector space. Such a model was introduced by Schütze [33, 30, 31]. It focusses on another aspect of human cognition that has been of particular interest to connectionist researchers: acquisition. Whereas other models take the senses of a word and the representation of semantic information for granted, the *context-group discrimination* model derives meaning representations and senses from corpora and provides a model of the acquisition of disambiguation information as well as a model of the disambiguation process itself. Lund et al. [20] have also argued for the use of such corpus-derived meaning representations. They show that findings about priming can be explained by vector representations which are derived from associations in text corpora. The following section describes context-group discrimination.

1.2 Context-Group Discrimination

In context-group discrimination, words, context and senses are represented in a multi-dimensional real-valued vector space as motivated by the previously cited research. There are two algorithms, one for sense acquisition, and one for discrimination proper.

In the acquisition phase, the occurrences of the ambiguous word in a training corpus are collected, their contexts are represented as context vectors and clustered using group average agglomerative clustering [37, 4]. The similarity measure for clustering is based on the neighbors of words in the corpus. The more neighbors two words have in common, the more similar they are. The more similar words appear in two contexts, the more similar the two contexts are. Occurrences are then grouped such that occurrences with similar contexts are assigned to the same cluster. Clusters are represented by their centroids (the average of their elements). In the application phase, the neighbor-based representation of the relevant context is computed and assigned to the cluster whose centroid it is closest to.

We will not use a connectionist implementation of clustering directly, but the Buckshot algorithm employed here is similar to connectionist clustering algorithms such as the ones described in [12] (Chapter 9 on "Unsupervised Competitive Learning").

The basic idea of context group discrimination is to induce senses from contextual similarity. There is some evidence that contextual similarity also plays a crucial role in human semantic categorization. In particular, Miller and Charles found evidence in several psychological experiments that humans determine the semantic similarity of words from the similarity of the contexts they are used in [23]. We hypothesize that, by extension, senses are also based on contextual similarity: a sense is a group of contextually similar occurrences of a word.

There are three types of entities in context-group discrimination: words, contexts, and senses. They are represented as word vectors, context vectors, and sense vectors, respectively. Word vectors are derived from neighbors in the corpus, context vectors are derived from word vectors, and sense vectors are derived by way of clustering from the distribution of context vectors.

1.2.1 Word Vectors

A vector for word w is derived from the close neighbors of w in the corpus. Close neighbors are all words that co-occur with w in a context of size k (i.e., a sliding window of k words centered around w). In the simplest case, the vector has an entry for each word that occurs in the corpus. The entry for word v in the vector for w records the number of times that word v occurs close to w in the corpus. Word vectors are formed by collecting a n-by-n matrix C, such that element c_{ij} records the number of times that words i and j cooccur in a window of size k. Column l (or equivalently row l) of matrix C represents word l. Note that C is symmetric since the words that are represented as word vectors are also those that form the dimensions of the n-dimensional space.

This vector representation captures the typical topic or subject matter of a word. For example, the vector of "astronaut" will have high counts for "spacecraft", "moon", and "NASA", indicating that the word "astronaut" is about the topic "outer space and its exploration". By looking at the amount of overlap between two vectors, one can roughly determine how closely they are related semantically. This is because related meanings are often expressed by similar sets of words. Semantically related words will therefore co-occur with similar neighbors and their vectors will have a lot of overlap.

This similarity can be measured by the cosine between two vectors. The value of the cosine is higher the more overlap there is between the neighbors of the two words whose vectors are compared. If two words occur with exactly the same neighbors (perfect overlap), then the value of the cosine is 1.0. In contrast, if there is no overlap at all, then the value of the cosine is 0.0. The cosine can therefore be used as a rough measure of semantic relatedness between words.

1.2.2 Context Vectors

The representation for words derived above conflates senses. For example, both senses of the word "suit" ("lawsuit" and "garment") are summed in its word vector. We need to go back to individual contexts in the corpus in order to acquire information about sense distinctions. We use *context vectors* to represent contexts. A context vector is the centroid (or sum) of the vectors of the words occurring in the context, weighted by the weights a_i (see below):

$$\vec{c} = \sum_i a_i \vec{v}(w_i)$$

where $\vec{c}$ is the context vector, the words w_i are the words occurring in the context and $\vec{v}(w_i)$ is the word vector of word w_i. Only words selected according to frequency or χ^2 are considered. See [39], [6], [24] for related algorithms for computing context vectors.

The centroid "averages" the direction of a set of vectors. For example, the centroid of two vectors is the vector that runs half way between them. The intuition underlying the computation of context vectors is that word vectors have several components, corresponding to different topics like "agriculture", "education", or "automobiles". If many of the words in a context have a strong component for one of the topics, say "automobiles", then the average of all vectors, the context vector, will also have a strong component for the topic. On the other hand, if only one or two words represent a particular topic, then the context vector will be weak on this component. So the context vector represents the strength of different topical or semantic components in a context.

In the computation of the context vector, a word vector is weighted according to its discriminating potential. A rough measure of how well word w_i discriminates between different topics is the log inverse document frequency used in information retrieval: [28]

$$a_i = \log(\frac{N}{n_i})$$

where n_i is the number of documents that w_i occurs in and N is the total number of documents.

Other algorithms for computing context vectors have been proposed in [39] (based on dictionary entries), [6] (based on hand-encoded semantic features), [7] (based on light parsing) and [24] (a comparison of dictionary-based and corpus-based context vectors).

1.2.3 Sense Vectors

Sense representations are computed as groups of similar contexts. All contexts of the ambiguous word are collected from the corpus. For each context, a context vector is computed. This set of context vectors is then clustered into a predetermined number of coherent clusters or context groups using group average agglomerative clustering [14]. The representation of a sense is simply the centroid of its cluster. It marks the portion of the multi-dimensional space which is occupied by the cluster.

1.2.4 Application of Context Group Discrimination

Context group discrimination uses word vectors and sense vectors as follows to discriminate occurrences of the ambiguous word. For an occurrence t of the ambiguous word v:

- Collect all selected words w_i that occur in t's context.
- For each w_i, retrieve its word vector $\vec{u}_i$.
- Compute the context vector $\vec{c}$ of t's context as the weighted sum of all word vectors $\vec{u}_i$.
- Retrieve all sense vectors $\vec{s}_j$ of v.
- Assign t to the sense j whose sense vector $\vec{s}_j$ is closest to $\vec{c}$.

This algorithm selects the context group whose sense vector is closest to the context vector of the occurrence of the word that is to be disambiguated. Context vectors and sense vectors capture semantic characteristics of the corresponding context and sense, respectively. Consequently, the sense vector that is closest to the context vector has the best semantic match with the context. Therefore, context-group discrimination categorizes the occurrence as belonging to that sense.

Note that the algorithm performs *discrimination* without *labeling*. It only tells us which occurrences belong to the same sense, not what the sense actually is. This is the reason that we call the algorithm context-group *discrimination*, as opposed to context-group disambiguation. Most other disambiguation algorithms do labeling in addition to discrimination. However, discrimination algorithms have two advantages. First, they need less training information since we do not need sense labels. Secondly, they can be argued to be cognitively more plausible since children also learn to discriminate between senses without having access to explicit labels.

capital	1	stock of goods
	2	seat of government
interest	1	a feeling of special attention
	2	a charge for borrowed money
motion	1	movement
	2	a proposal for action
plant	1	a factory
	2	living being
ruling	1	an authoritative decision
	2	to exert control, or influence
space	1	area, volume
	2	outer space
suit	1	an action or process in a court
	2	a set of garments
tank	1	a combat vehicle
	2	a receptacle for liquids
train	1	a line of railroad cars
	2	to teach
vessel	1	a ship or plane
	2	a blood vessel

Table 1.1: Senses of ambiguous words.

capital/s	13015	200	2%	64
interest/s	21374	200	4%	58
motion/s	2705	200	0%	55
plant/s	12833	200	0%	54
ruling	5482	200	3.5%	60
space	9136	200	0%	56
suit/s	7467	200	12.5%	57
tank/s	3909	200	4.5%	90
train/s	4271	200	1.5%	74
vessel/s	1618	144	13.9%	69

Table 1.2: Number of occurrences of test words in training and test set, percent rare senses in test set, and baseline performance (all occurrences assigned to most frequent sense).

Tables 1.1, 1.2, and 1.3 evaluate the performance of the algorithm on ten ambiguous words [32]. Table 1.1 defines the senses of the words. Table 1.2 gives the number of occurrences of the ambiguous words in training set and test set and the baseline of disambiguation performance (i.e., always choosing the most frequent sense). In Table 1.3, the performance of the algorithm is

number of clusters:	2	10
capital	93	90
interest	92	79
motion	89	72
plant	80	94
ruling	90	76
space	66	71
suit	94	94
tank	90	90
train	83	75
vessel	69	94
average	87	88

Table 1.3: Results of two disambiguation experiments (2 and 10 clusters per word).

measured by the percentage of correct disambiguation decisions, for 2 clusters and for 10 clusters. The algorithm performs well although somewhat worse than methods that derive sense definitions from thesauri [40] or dictionaries [41].

1.3 A Large-Scale Application: Information Retrieval

Some connectionist researchers see connectionism as a subfield of statistics and emphasize the statistical interpretation of connectionist models [25, 26]. As an endeavor in statistical learning, connectionism is concerned with a particular class of learning models, neural networks. Neural networks have a number of properties that distinguish them from other forms of statistical learning. These properties are motivated by learning needs that arise in biological systems. For example, neural networks provide shared structure for tackling several learning problems at the same time, for example the visual recognition of different types of objects. The transfer of knowledge from one learning task to another is characteristic of learning in animals and humans. Neural networks are also trained pattern by pattern, i.e., by presenting one pattern at a time and adjusting the parameters of the network such that the response to this particular pattern is closer to the desired outcome. Again, incremental learning is more plausible for biological systems than "batch" learning since realistic learning only has direct access to one stimulus (visual, auditory or otherwise) at a time.

In this section, we will concentrate on a third interest of connectionist researchers in the realm of statistical learning: the ability to handle large data sets. Many neural network architectures fit standard statistical models. For example, a neural network with no hidden layers, one output node and a sigmoid activiation function corresponds to the well-known logistic model. However, connectionists pay particular attention to problems that occur when models are trained on large data sets. On-line learning and gradient descent algorithms can accommodate data sets that are difficult to handle with traditional statistical fitting methods. In addition, neural network architectures can be implemented in a highly parallel fashion, which also results in faster computation.

In the connectionist tradition of research, we will now summarize the information retrieval experiment described in [34], in which the context-group discrimination algorithm exploits the systematicities of word sense occurrences in large text corpora. Information retrieval is an application where both acquisition of senses and efficiency are important. The acquisition of senses from the corpus of application is important because general dictionaries often do not cover senses of technical terms. For example, the financial sense of "distribution" in "mutual fund distributions" may not be covered in a general dictionary, whereas it can be induced from a text corpus if there are enough usages of the word.

Efficiency is equally important in information retrieval (IR) applications. If word sense disambiguation is to be of real help in information retrieval, then it has to be applicable to the large and very large collections (like the World Wide Web) that are typical today. We will report here on an experiment on a medium-size test collection, category B of the evaluation in the TREC-1 IR system comparison [9]. The corpus has roughly 170,000 documents and 500 megabytes of text from the Wall Street Journal and a set of 25 evaluation queries (Topics 51-75).

	word-based	sense-based		combined	
at 0.00	0.693	0.788	+13.7	0.854	+23.2
at 0.10	0.540	0.629	+16.5	0.645	+19.4
at 0.20	0.453	0.503	+11.0	0.506	+11.7
at 0.30	0.385	0.427	+10.9	0.434	+12.7
at 0.40	0.315	0.340	+7.9	0.347	+10.2
at 0.50	0.264	0.260	-1.5	0.291	+10.2
at 0.60	0.206	0.198	-3.9	0.229	+11.2
at 0.70	0.165	0.145	-12.1	0.174	+5.5
at 0.80	0.118	0.110	-6.8	0.129	+9.3
at 0.90	0.085	0.076	-10.6	0.091	+7.1
at 1.00	0.061	0.057	-6.6	0.059	-3.3
average precision (non-interpolated) over all rel docs					
	0.299	0.321	+7.4	0.342	+14.4

Table 1.4: Disambiguation markedly improves retrieval performance.

The sense discrimination algorithm described above can be applied directly to this test colleciton by replacing the words in a standard "bag of words" text representation by word senses. That is, the text is analyzed into words and each word occurrence annotated by a sense as suggested by the discrimination algorithm. These annotated, and hence differentiated, words may then be used as the basic features for retrieval matching and scoring.

The queries contain 1013 different word types (excluding stop words). We divided each term into as many senses as could be reliably distinguished from the available data. Our experience indicates that 50 occurrences per sense is a reasonable estimate of the minimum amount of data necessary to distinguish between senses. Therefore, the number of senses for word w was defined to be $\frac{f}{50}$ where f is the number of occurrences of w in the corpus. Only 20 senses were considered even for words that occurred more often than 1000 times. In the experiment, the occurrences of each of the 1013 query terms in the corpus were clustered into the predetermined number of clusters.

For disambiguation, context vectors for each occurrence of one of the query terms in the corpus were computed and assigned to the closest sense vector. Each document in the collection was indexed with the automatically assigned senses rather than the terms themselves. Finally, the 25 queries were processed in the same way.

A modification of the standard vector similarity model that represents documents and queries as vectors [29] was then applied for senses instead of words. In the standard model, documents are ranked according to the number of words they share with the query. In the modified model, documents are ranked according to the number of senses (disambiguated words) they share with the query. We used term frequency/inverse document frequency weighting [28] for the senses as well as for the words. We achieved a measurable improvement for sense-based retrieval when compared to word-based retrieval: Average precision for 11 points of recall increased by 14.4%.

The improvement in retrieval performance shows that the adaptibility of our word sense discrimination algorithm (i.e., its ability to acquire senses from text) and its efficiency make it a good match for information retrieval applications.

1.4 Discussion

In this chapter, we presented a word sense disambiguation algorithm inspired by concerns of connectionist research like large-scale applicability and psychological plausibility. The importance of the concept of activation for explaining human word sense disambiguation was the motivation for building a mechanical disambiguation model in which words, contexts and senses are represented as vectors of activations in a multi-dimensional, real-valued vector space. The connectionist interest in extracting regularites from large data sets leads to the consideration of a large-scale application of disambiguation to information retrieval.

The word sense disambiguation algorithm presented in this chapter partly draws its inspiration from connectionism's emphasis on acquisition. Other disambiguation algorithms employ different sources of information: hand-constructed rules [16], online dictionaries [19], [17], [8], knowledge bases [13], bilingual corpora [1], [2], bilingual dictionaries [5], hand-labeled training sets [11], [18],

[24] and thesauri [38], [40].

What distinguishes context group discrimination from other work on disambiguation is that no outside source of information need be supplied as input to the algorithm. While it is no realistic model of human acquisition of senses, it at least begins to address the question of how senses might be acquired. The algorithm relies on contextual similarity as the key relation to form senses.

The number of senses of a particular word is a parameter of context-group discrimination, that is, it needs to be determined in advance by some external criterion. This criterion was the number of occurrences in the IR application. This aspect of context-group discrimination is often met with skepticism. However, it is becoming increasingly obvious that there is not one single number of correct senses for a particular word. The level of individuation of senses, fine or coarse, ultimately determines what the right number of senses is. A novel feature of our IR application is that we do not choose the same level of individuation for all words. Instead, a finer level of sense individuation is chosen for frequent words than for rare words since there is more evidence for the former to make fine sense distinctions reliably.

There is some evidence that contextual similarity plays a crucial role in human semantic categorization. In particular, Miller and Charles found evidence in several psychological experiments that humans determine the semantic similarity of words from the similarity of the contexts they are used in. They summarized this result in the following hypothesis.

> *Strong Contextual Hypothesis:* Two words are semantically similar to the extent that their contextual representations are similar. [23]

A contextual representation of a word is knowledge of how that word is used. So the hypothesis states that semantic similarity is determined by how similar the sets of contexts are that the two words can be used in.

The hypothesis that underlies context group discrimination is an extension of the Strong Contextual Hypothesis to senses:

> *Contextual Hypothesis for Senses:* Two occurrences of an ambiguous word belong to the same sense to the extent that their contextual representations are similar.

So a sense is simply a group of occurrences with similar contexts. The analogy between the contextual hypotheses for words and senses is that both word types and word occurrences are semantically similar to the extent that their contexts are semantically similar. A group of contextually similar word occurrences is a sense. Miller and Charles' work thus provides a justification for our framework, the induction of senses from contextual similarity.

Bibliography

[1] Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. Word-sense disambiguation using statistical methods. In *Proceedings of ACL 29*, pages 264–270, Morristown NJ, 1991. Association of Computational Linguistics.

[2] Kenneth W. Church and William A. Gale. Concordances for parallel text. In *Proceedings of the Seventh Annual Conference of the UW Centre for the New OED and Text Research*, pages 40–62, England: Oxford, 1991.

[3] Garrison W. Cottrell. A model of lexical access of ambiguous words. In Steven L. Small, Garrison W. Cottrell, and Michael K. Tanenhaus, editors, *Lexical Ambiguity Resolution: Perspectives from Psycholinguistics, Neuropsychology, and Artificial Intelligence*, pages 179–194. Morgan Kaufmann, San Mateo CA, 1988.

[4] Douglas R. Cutting, Jan O. Pedersen, David Karger, and John W. Tukey. Scatter/gather: A cluster-based approach to browsing large document collections. In *Proceedings of SIGIR '92*, pages 318–329, New York, 1992. Association of Computing Machinery.

[5] Ido Dagan, Alon Itai, and Ulrike Schwall. Two languages are more informative than one. In *Proceedings of ACL 29*, pages 130–137, Morristown NJ, 1991. Association of Computational Linguistics.

[6] Stephen I. Gallant. A practical approach for representing context and for performing word sense disambiguation using neural networks. *Neural Computation*, 3(3):293–309, 1991.

[7] Gregory Grefenstette. *Explorations in Automatic Thesaurus Discovery.* Kluwer Academic Press, Boston, 1994.

[8] Joe A. Guthrie, Louise Guthrie, Yorick Wilks, and Homa Aidinejad. Subject-dependent co-occurrence and word sense disambiguation. In *Proceedings of ACL 29*, pages 146–152, Morristown NJ, 1991. Association of Computational Linguistics.

[9] Donna Harman. Overview of the first trec conference. In *Proceedings of SIGIR '93*, pages 36–47, 1993.

[10] Catherine L. Harris. Connectionism and cognitive linguistics. *Connection Science*, 2:7–33, 1990.

[11] Marti A. Hearst. Noun homograph disambiguation using local context in large text corpora. In *Seventh Annual Conference of the UW Centre for the New OED and Text Research*, pages 1–22, England: Oxford, 1991.

[12] John A. Hertz, Richard G. Palmer, and Anders S. Krogh. *Introduction to the theory of neural computation.* Addison-Wesley, Redwood City CA, 1991.

[13] Graeme Hirst. *Semantic Interpretation and the Resolution of Ambiguity.* Cambridge University Press, Cambridge, 1987.

[14] Anil K. Jain and Richard C. Dubes. *Algorithms for Clustering Data.* Prentice Hall, Englewood Cliffs NJ, 1988.

[15] Alan H. Kawamoto. Distributed representations of ambiguous words and their resolution in a connectionist network. In Steven L. Small, Garrison W. Cottrell, and Michael K. Tanenhaus, editors, *Lexical Ambiguity Resolution: Perspectives from Psycholinguistics, Neuropsychology, and Artificial Intelligence*, pages 195–228. Morgan Kaufmann, San Mateo CA, 1988.

[16] Edward Kelly and Phillip Stone. *Computer Recognition of English Word Senses*. North-Holland, Amsterdam, 1975.

[17] Robert Krovetz and W. Bruce Croft. Word sense disambiguation using machine-readable dictionaries. In *Proceedings of SIGIR '89*, pages 127–136, New York, 1989. Association of Computing Machinery.

[18] Claudia Leacock, Geoffrey Towell, and Ellen Voorhees. Corpus-based statistical sense resolution. In *Proceedings of the ARPA Workshop on Human Language Technology*, San Mateo CA, 1993. Morgan Kaufman.

[19] Michael Lesk. Automatic sense disambiguation: How to tell a pine cone from an ice cream cone. In *Proceedings of the 1986 SIGDOC Conference*, pages 24–26, New York, 1986. Association for Computing Machinery.

[20] Kevin Lund, Curt Burgess, and Ruth Ann Atchley. Semantic and associative priming in high-dimensional semantic space. In J. D. Moore and J. F. Lehman, editors, *Proceedings of the Seventeenth Annual Meeting of the Cognitive Science Society*, Hillsdale NJ, 1995. Erlbaum Associates.

[21] J. L. McClelland, Mark St. John, and Roman Taraban. Sentence comprehension: A parallel distributed processing approach. *Language and cognitive processes*, 4(3/4 SI):287–335, 1989.

[22] James L. McClelland, David E. Rumelhart, and the PDP Research Group, editors. *Parallel Distributed Processing. Explorations in the Microstructure of Cognition. Volume 2: Psychological and Biological Models.* The MIT Press, Cambridge MA, 1986.

[23] George A. Miller and Walter G. Charles. Contextual correlates of semantic similarity. *Language and Cognitive Processes*, 6(1):1–28, 1991.

[24] Yoshiki Niwa and Yoshihiko Nitta. Co-occurrence vectors from corpora vs. distance vectors from dictionaries. In *Proceedings of COLING94*, pages 304–309, 1994.

[25] Michael Richard and Richard Lippman. Neural network classifiers estimate bayesian a posteriori probabilities. *Neural Computation*, 3:461–483, 1991.

[26] David E. Rumelhart, Richard Durbin, Richard Golden, and Yves Chauvin. Backpropagation: The basic theory. In Yves Chauvin and David E. Rumelhart, editors, *Back-propagation: Theory, Architectures, and Applications*, pages 1–34. Lawrence Erlbaum, Hillsdale NJ, 1995.

[27] David E. Rumelhart, James L. McClelland, and the PDP research group, editors. *Parallel Distributed Processing. Explorations in the Microstructure of Cognition. Volume 1: Foundations.* The MIT Press, Cambridge MA, 1986.

[28] Gerard Salton and Chris Buckley. Improving retrieval performance by relevance feedback. *Journal of the American Society for Information Science*, 41(4):288–297, 1990.

[29] Gerard Salton and Michael J. McGill. *Introduction to modern information retrieval.* McGraw-Hill, New York, 1983.

[30] Hinrich Schütze. Dimensions of meaning. In *Proceedings of Supercomputing '92*, pages 787–796, Los Alamitos CA, 1992. IEEE Computer Society Press.

[31] Hinrich Schütze. Word space. In Stephen José Hanson, Jack D. Cowan, and C. Lee Giles, editors, *Advances in Neural Information Processing Systems 5*, pages 895–902. Morgan Kaufmann Publishers, San Mateo CA, 1993.

[32] Hinrich Schütze. *Ambiguity Resolution in Language Learning.* CSLI Publications, Stanford CA, 1997.

[33] Hinrich Schütze. Automatic word sense discrimination. *Computational Linguistics*, 24(1):97–124, 1998.

[34] Hinrich Schütze and Jan O. Pedersen. Information retrieval based on word senses. In *Fourth Annual Symposium on Document Analysis and Information Retrieval*, pages 161–175, Las Vegas NV, 1995.

[35] David A. Swinney. Lexical access during sentence comprehension: (Re)Consideration of context effects. *Journal of Verbal Learning and Verbal Behavior*, 18:645–659, 1979.

[36] Michael K. Tanenhaus, James M. Leiman, and Mark S. Seidenberg. Evidence for multiple stages in the processing of ambiguous words in syntactic contexts. *Journal of Verbal Learning and Verbal Behavior*, 18:427–440, 1979.

[37] C. J. van Rijsbergen. *Information Retrieval.* Butterworths, London, 1979. Second Edition.

[38] Donald E. Walker and Robert A. Amsler. The use of machine-readable dictionaries in sublanguage analysis. In Ralph Grishman and Richard Kittredge, editors, *Analyzing language in restricted domains: sublanguage description and processing*, pages 69–84. L. Erlbaum Associates, Hillsdale NJ, 1986.

[39] Yorick A. Wilks, Dan C. Fass, Cheng ming Guo, James E. McDonald, Tony Plate, and Brian M. Slator. Providing machine tractable dictionary tools. *Journal of Computers and Translation*, 2, 1990.

[40] David Yarowsky. Word-sense disambiguation using statistical models of Roget's categories trained on large corpora. In *Proceedings of Coling-92*, pages 454–460. Association of Computational Linguistics, 1992.

[41] David Yarowsky. Unsupervised word sense disambiguation rivaling supervised methods. In *Proceedings of ACL 33*, 1995.