Self-Organizing Maps of Words for Natural Language Processing Applications

Timo Honkela
Neural Networks Research Centre
Helsinki University of Technology
P.O.Box 2200, FIN-02015 HUT, Finland
http://nucleus.hut.fi/~tho/

Keywords: self-organizing map, neural networks, semantics, natural language processing, full-text information retrieval

Abstract. Kohonen's Self-Organizing Map (SOM) is a general unsupervised tool for ordering high-dimensional statistical data so that neighboring nodes on the map represent similar inputs. Often the SOM is applied to numerical data in application areas like pattern recognition, signal processing, and multivariate statistical analysis. The SOM can also be used to find statistical similarities between symbols if suitable contextual information is available as there indeed is for words in natural language texts. Using the short contexts of each word occurring in a text the SOM algorithm is able to organize the words into grammatical and semantic categories represented on a two-dimensional array. The similarity of the categories is reflected in their distance relationships on the array. This kind of a word category map may then be utilized in applications such as the analysis of large document collections. In addition to the description of information retrieval and textual data mining applications, this paper outlines the relation of word category maps to the symbolic knowledge representation formalisms. The graded nature of the categorization performed by the SOM is discussed. The aim is also to provide an overview on the research results and on the potential new areas.

1. Introduction

Development of large-scale natural language processing applications is restricted by quantitative and qualitative limitations. Quantitatively, a practical system requires a substantial knowledge base even if it is intended to be used in a moderately narrow domain. One solution for this problem has been an approach where vast common repositories of knowledge items (frames, facts, rules) have been collected. Qualitatively problematic areas remain, though, e.g., graded phenomena, inherent ambiguity of natural language, and subjectivity and variation in natural language generation and interpretation. Gradual changes in the domain of the application make non-adaptive systems vulnerable.

In this paper, the problem of characterizing the semantics of words, or more specifically lexical elements, is considered. The approach is data-driven: large collections of input texts are analyzed using Self-Organizing Maps (SOMs) (Kohonen 1982, 1995). The SOM both projects a data set into a more illustrative form on a lower-dimensional space and reduces the number of data items into a smaller number of map locations. The representations become ordered according to their similarity relations in an unsupervised learning process. This organizing property makes it especially useful for analyzing and visualizing large data collections.

2. Word category maps

Word category maps (also called self-organizing semantic maps; see Ritter and Kohonen 1989, 1990) are SOMs that have been organized according to word similarities, measured by the similarity of the short contexts of the words.

The SOM algorithm is based on competitive learning: the artificial neurons of the network gradually become sensitive to different input categories. The neurons form a regular usually two-dimensional array, i.e., a map. When an input vector x is processed, the best-matching neuron on the map, i.e., the neuron that is closest in the (normally Euclidean) metric wins the competition and it is updated as well as its neighborhood according to the adaptation rule (Kohonen 1995):

$$m_i(t+1) = m_i(t) + h_{ci}(t)[x(t) - m_i(t)]$$

Here $x(t)$ is the input vector at time t and $c = c(x(t))$ is the index of the winning neuron. A model vector m_i is associated with each neuron. The vector m_i specifies the coordinates of a neuron in the original input space. The amount the winning unit and its neighbors adapt is governed by the neighborhood kernel $h_{ci}(t)$. The kernel is chosen to be wide in the beginning of the learning process to facilitate global ordering of the map. The width and the height of the kernel decrease slowly during the learning. For a detailed description of the SOM algorithm, its mathematical analysis and multitude of application examples see (Kohonen 1995).

In the unsupervised formation of the word category map, each input x consists of an encoded word and its averaged context. Each word in the vocabulary is encoded as a n-dimensional random vector. In our experiments (e.g., Honkela et al. 1995, Kaski et al. 1996) n has usually been 90. A more straightforward way to encode each word would be to reserve one component for each distinct word but then, especially with a large vocabulary, the dimensionality of the vector would be computationally intractable. A mathematical analysis of the dimensionality reduction based on random encoding is presented by Ritter and Kohonen (1989).

The basic steps in forming a word category map are listed in the following:

1. Create a unique random vector for each word in the vocabulary.
2. Find all the instances of each word to be mapped in the input text collection. Such words will be called key words in the following. Calculate the average over the contexts of each key word. The random codes formed in step 1 are used in the calculation. The context may consist of, e.g., the preceding and the succeeding word, or some other window over the context. As a result each key word is associated with a contextual fingerprint. If the original random encoding for a single word is, for instance, 90-dimensional, the resulting vector of key word with context is 270-dimensional if one neighboring word from both sides is included. The classifying information is in the context. Therefore the key word part of the vector is multiplied by a small scalar, ε. In our experiments ε has usually been 0.2.
3. Each vector formed in step 2 is input to the SOM. The resulting map is labeled after the training process by finding best-matching neuron for each vector and labeling the neuron with the key part of the vector.

Fig. 1 outlines the automatic partition of the words into syntactic classes (see Honkela et al. 1995 for details of the experiment). The input for the map was the English translation of Grimm fairy tales. In the resulting map the verbs form an area of their own in the top of the map whereas the nouns can be found in the opposite corner. The modal verbs are in one area. Some semantically oriented ordering can also be found: for instance, the inanimate and animate nouns form separate clusters. Conceptually interrelated words tend to fall into the same or neighboring map nodes (see Fig. 2 and Kohonen 1996b for the details of the experiment). Nodes may thus be viewed as word categories. The overall organization of a word category map reflects the syntactic categorization of the words. The SOM performs a non-linear mapping from the original multi-dimensional space into the two-dimensional lattice. To visualize the non-linearities, one may use U-matrices (Ultsch 1993) to show the distances between neighboring map nodes in the original input space. In Fig. 2 the shades of gray are used in this task : dark areas denote to the relatively larger distances between the light clusters.

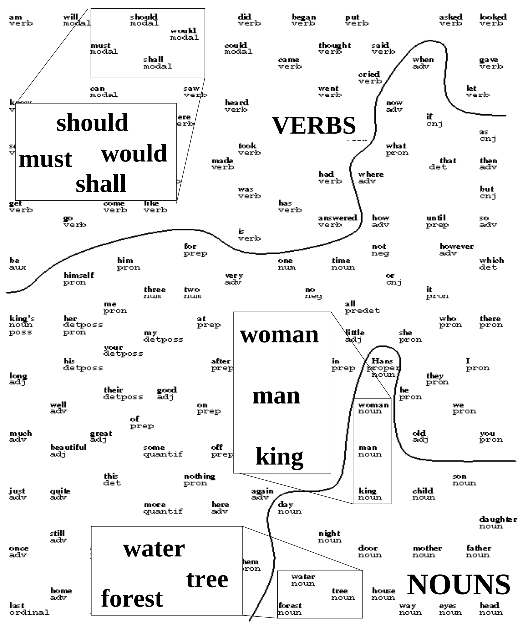

Figure 1. An example of organization of 150 most common words of Grimm fairy tales into a two-dimensional map. The apparent category areas have been automatically formed using SOM (Honkela et al. 1995).

For more detailed information on the word category map formation and examples see, for instance, (Ritter and Kohonen 1989, 1990; Honkela et al, 1995). Several other studies have also been published on SOMs that map words into grammatical and semantic categories by, for instance, Finch and Chater (1992), Miikkulainen (1993), and Scholtes (1993).

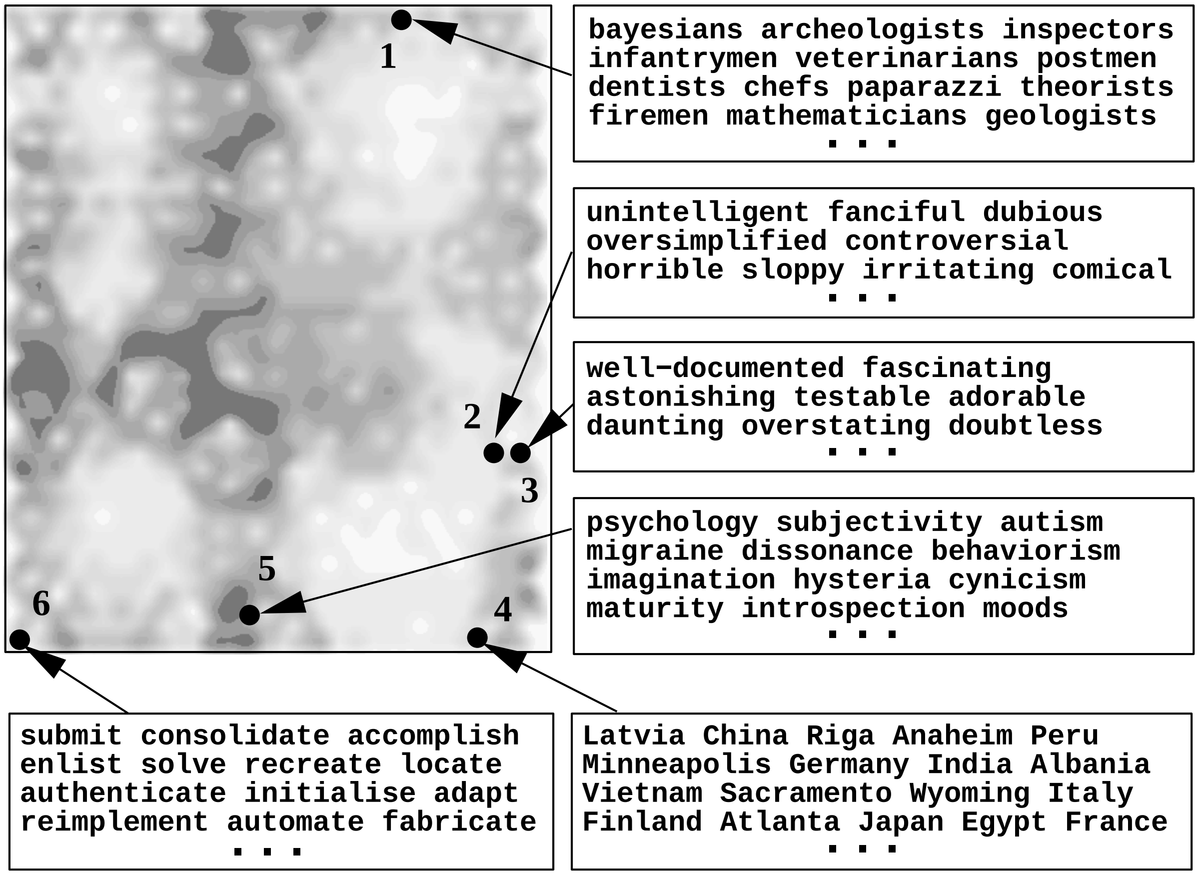

Figure 2. Some examples of node contents in a 24 times 32 word category map. The input text consisted of a large collection of Usenet newsgroup articles. Only a selection of the content of each node is visualized.

In the following, the relation between the word category maps and the representational formalisms of symbolic AI such as semantic nets is discussed. Practical application areas of the maps are outlined with special consideration of information retrieval and textual data mining. Finally, a general approach of interpreting the responses of a word category map as values of fuzzy category membership values is presented. Arguments for regarding word category maps as "true semantic networks" are given.

3. Word category maps vs. semantic nets

Models developed in artificial intelligence have widely been used as explanatory tools in cognitive science. Semantic networks are central to the field, often suggested to provide a model of the structure and operations of the memory. In a semantic net both the nodes and the arcs are labeled. In the modeling view, a linked pair of nodes can be considered as a propositional structure.

In semantic networks the meaning of the nodes and the links are determined by a human. Some inductive inference methods may be used to make generalizations but they are based on the static conceptual structure determined by the network. The staticness holds also for other symbolic knowledge representation formalisms such as frame systems. The reference relations of the nodes in a SOM, however, emerge through an adaptive process. As a result the relation between the network state and the portion of "external reality" presented for the network is not directly determined by a human being. Especially the unsupervised learning paradigm that is in use in SOM provides means for autonomous status of the network. The adaptive process is, of course, dependent on the training material but the interpretations or classification are not given by a human as they are in the supervised learning systems such as the backpropagation of error algorithm. For further points of view see (Ritter and Kohonen 1989) for an insightful multidisciplinary discussion on the nature of categorization.

One may ask what is the motivation behind using the symbol part in the input vector rather than plainly labeling the map afterwards. In the basic applications also the latter approach may be valid but it excludes many useful possibilities such as further augmenting the vector with graded memberships, having multiple symbols with varying weights included, and enhancing the representation with non-symbolic classifying information. A model in which the symbol grounding is based on a simultaneous exposition of the symbol and the context seems also more natural as a cognitive model than the approach in which labeling is accomplished afterwards.

The comparison between static symbolic knowledge representation and self-organization of numerical representation was presented here in order to highlight some of the fundamental differences between these approaches. The differences also influence the nature of the practical applications.

4. Applications of word category maps

Word category maps can be used in practical large-scale natural language processing applications such as intelligent information retrieval and textual data mining.

4.1. Maps of documents: WEBSOM method

With the aid of the word category map any document can be encoded as a word category histogram. Closely related words that are in the same category on the map then contribute identically to the code formed for a document. In the WEBSOM method developed in our laboratory (see, e.g., Honkela et al. 1996a, 1996b, Kaski et al. 1996, Kaski 1997a, 1997b, Kohonen et al. 1996b, Lagus 1996; also the demo in the WWW address http://websom.hut.fi/websom/) the Self-Organizing Map (SOM) algorithm is used to order documents based on their full textual contents. The ordering of the documents is achieved by a two-level analysis. First the word categories are extracted from the text. In the second phase the location of each document on the document map is determined. The text of a document is mapped onto the word category map whereby a histogram of the "hits" on it is formed. To reduce the sensitivity of the histogram to small variations in the document content, the histograms are smoothed. The document map is formed with the SOM algorithm using the histograms as ''fingerprints" of the documents. To speed up computation, the positions of the word labels on the word category map may be looked up by hash coding. Kaski (1997a) presents analyses of the WEBSOM method and other applications of the SOM in data exploration.

4.2. WEBSOM applications

The visualized order on a document map can be utilized for an explorative search or exploration of novel knowledge areas, whereby the scope can be changed interactively. The WEBSOM provides a content-based alternative to information retrieval based on keyword searches. A holistic two-dimensional view on large document collections can be acquired automatically. WEBSOM can also be used to solve the problems related to common variation in expression, i.e., different ways of expressing the same idea.

In principle the WEBSOM method is applicable to many kinds of text collections and thus it has a wide area of potential applications ranging from organizing an individual's own texts into creating a digital library of a large. Largest map published so far contained 131 500 full-text documents on a map with 12 288 nodes (Kohonen 1996b). The stochastic nature of the processing and the variability of the input material must be remembered when judging the results. In some cases external information may be needed while the text collection does not, as such, contain enough useful contextual information. For instance, a small document collection or a very heterogeneous collection may not be well suited for the WEBSOM method. Gallant (1991) and Hecht-Nielsen (1994) present the possibility to acquire the input material by hand so that each word to be encoded is evaluated by a human with respect to a limited number of features (perhaps 200 to 500). That kind of approach is, of course, less ambitious than that of WEBSOM.

4.3. Other document maps and application areas

The SOM has previously been utilized to form document maps, e.g., by Lin et al (1991), Scholtes (1991a, 1991b, 1993), Merkl (1993, 1995), and Chen et al. (1996). The novel feature in WEBSOM in comparison with the previous studies is the two-level processing that includes an initial analysis by a word category map. In addition to using words, Scholtes has successfully used character n-grams in document encoding. He has also used SOM as a "black box" in finding texts of interest from an incoming document flow. Chen et al. describe several examples of using a multi-layered SOM architecture to, e.g., create large maps of Internet home pages. In addition to the methodological and technical description they discuss the results of the user evaluation of their system. Merkl has used document maps to facilitate software library organization based on the textual descriptions of the software components. Streeter and Lochbaum (1988) show one potential application area of word category maps by applying singular value decomposition applied to term-document matrix: characterizing organizations based on their technical documentation.

When looking at an ordered map like the one in Fig 2, it becomes obvious that word category maps may be helpful in constructing thesauri. Also other areas of natural language processing such as machine translation require a lot of lexical information. In machine translation one central problem is the lexical selection between multiple word or phrase alternatives in the generation phase. A SOM-based contextual processing system may provide a useful solution for this task.

5. On category membership

The SOM has been shown to be able to automatically create implicit emergent categories from uncategorized linguistic input. The map nodes may be considered as adaptive prototypes (Honkela 1997). The SOM gives a graded response when an input is positioned on a map. That provides a potential solution for acquiring fuzzy membership values in a multidimensional feature space (see also Mitra and Pal 1994). If one requires that, for example, tallness is defined with respect to some parameters such as length of a person, sex and age, one can input positive, negative and neutral (unclassified) examples containing the parameter values and the symbol part that in this case consists of only one vector element. Here the neutral example refers to a case in which the symbol part is missing. Otherwise the learning may be considered as self-supervised. Fig. 3 shows the results of a small experiment where 200 examples were given as the input.

In this example the input dimensions were predetermined and they can be considered as symbolic. As the SOM is able to create conceptual clusters provided that suitable input material is available, one may foresee that it is also possible to create fuzzy categories using SOM without predetermined readily interpreted dimensions. It is important to notice here that the SOM does not require deciding the number of clusters beforehand. In practice this approach would mean that, for instance, the input would originate from a signal source (speech or pictorial images, etc.) and the SOM would find the relevant

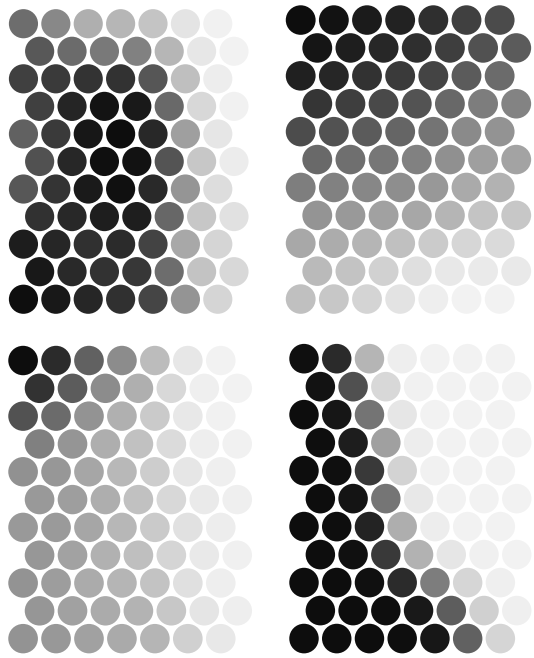

Figure 3. Component maps of an experiment of graded classification. The components are sex, age, height; and tallness starting from the upper left corner, and ending up to the lower right map of tallness. In the tallness component map the dark areas refer to values close to one. The result of this miniature experiment is included here only to provide a concrete example related to the principles discussed in the text.

clusters that are associated with the linguistic expression. The SOM would be the central component of such system but naturally other tools for preprocessing would also be needed. This kind of possibility must be regarded here as a potential research direction and a conceptual or philosophical framework rather than a ready system or method. Such an approach "only" requires, though, the unsupervised adaptive capability of the SOM and abandoning the symbolic flavors of the processing altogether: the input components would not have a prior interpretation -- the interpretations would emerge through self-organization!

6. Conclusions

In this paper some application areas of self-organizing maps of words have been described. The most popular application of Self-Organizing Maps (SOMs) among natural language processing is currently information retrieval and textual data mining. The SOM has proven to be very useful in providing an overall view on large document collections. It can also be used in creating an associative search method so that one may position a new document on a map and find the closely related texts in the neighborhood on the document landscape. Thus SOM provides very useful means both for exploration and search. The other potential applications of word category maps include thesarus construction and lexical selection in machine translation.

The methodological relevance of word category maps has also been studied in this paper in comparison with static and symbolic methods for knowledge representation as well as when compared with methods based on supervised learning. The competitive learning approach of SOM provides a good basis to be used as a cognitive model in the area of natural language interpretation.

Acknowledgments

I wish to express my gratitude to Teuvo Kohonen, Samuel Kaski, and Krista Lagus.

References

Chen, H., Schuffels, C., and Orwig, R. (1996): Internet categorization and search: a self-organizing approach. In *Journal of Visual Communication and Image Representation*, vol. 7, no. 1, pp. 88-102.

Finch, S., and Chater, N. (1992): Unsupervised methods for finding linguistic categories. In I. Aleksander and J. Taylor (eds.), *Artificial Neural Networks*, vol. 2, pp. II-1365-1368, North-Holland.

Gallant, S.I. (1991): A practical approach for representing context and for performing word sense disambiguation using neural networks. In *Neural Computation*, 3, pp. 293-309.

Hecht-Nielsen, R. (1994): Context vectors - general purpose approximate meaning representations self-organized from raw data. In J.M. Zurada, R.J. Marks II, and C.J. Robinson (eds.), *Computational Intelligence*, IEEE Press, New York.

Honkela, T., Pulkki, V., and Kohonen, T.. (1995): Contextual relations of words in Grimm tales analyzed by self-organizing map. In F.Fogelman-Soulié and P.Gallinari (eds), *ICANN-95, Proceedings of International Conference on Artificial Neural Networks*, vol. 2, pp. 3-7. EC2 et Cie, Paris, 1995.

Honkela, T., Kaski, S., Lagus, K., and Kohonen, T. (1996a): *Newsgroup exploration with WEBSOM method and browsing interface*. Technical Report A32, Helsinki University of Technology, Laboratory of Computer and Information Science, Espoo, January 1996. WEBSOM home page (1996) available at http://websom.hut.fi/websom/.

Honkela, T., Kaski, S., Lagus, K., and Kohonen, T. (1996b): Exploration of full-text databases with Self-Organizing Maps. In *Proceedings of International Conference on Neural Networks (ICNN-96)*, vol. 1, pp. 56-61, IEEE, Piscataway, NJ.

Honkela, T. (1997): Emerging categories and adaptive prototypes: Self-organizing maps for cognitive linguistics. Extended abstract, accepted to the International Cognitive Linguistics Conference, Amsterdam, 14.-19.7.1997.

Kaski, S., Honkela, T., Lagus, K., and Kohonen, T. (1996): Creating an order in digital libraries with self-organizing maps. In *Proceedings of WCNN-96, World Congress on Neural Networks*,.pp. 814-817, Lawrence Erlbaum and INNS Press, MahWah, NJ.

Kaski, S. (1997a): *Data Exploration Using Self-Organizing Maps*. PhD Thesis. Helsinki Univ. of Technology, Department of Computer Science and Engineering, Acta Polytechnica Scandinavica, Mathematics, Computing and Management in Engineering Series, no 82.

Kaski, S. (1997b): Computationally efficient approximation of a probabilistic model for document represenation in the WEBSOM full-text analysis method. Accepted to *Neural Processing Letters*.

Kohonen, T. (1982): Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, 43, pp. 59-69.

Kohonen, T. (1995): *Self-Organizing Maps*. Springer-Verlag.

Kohonen, T., Hynninen, J., Kangas, J., and Laaksonen, J. (1996a): *SOM_PAK: The self-organizing map program package*. Technical Report A31, Helsinki University of Technology, Laboratory of Computer and Information Science, Espoo, Finland.

Kohonen, T., Kaski, S., Lagus, K., and Honkela, T. (1996b): Very large two-level SOM for the browsing of newsgroups. In *Proceedings of International Conference on Artificial Neural Networks*, Lecture Notes in Computer Science, vol. 1112, pp. 269-274, Springer, Berlin.

Lagus, K., Honkela, T., Kaski, S., and Kohonen, T. (1996): Self-organizing maps of document collections: A new approach to interactive exploration. In *Proceedings. of the 2nd International Conference on Knowledge Discovery and Data Mining (KDD-96)*, AAAI Press, Menlo Park, California.

Lin, X., Soergel, D., and Marchionini, G. (1991): A self-organizing semantic map for information retrieval. In *Proceedings of 14th Annual International ACM/SIGIR Conference on Research & Development in Information Retrieval*, pp. 262-269, 1991.

Merkl, D. (1993): Structuring software for reuse - the case of self-organizing maps. In *Proceedings of IJCNN-93-Nagoya, International Joint Conference on Neural Networks*, vol. III, pp. 2468-2471. IEEE Service Center. Piscataway, NJ, 1993.

Merkl, D. (1995): Content-based software classification by self-organization. *Proceedings of the IEEE International Conference on Neural Networks (ICNN'95),* Perth, Australia.

Miikkulainen, R. (1993): *Subsymbolic Natural Language Processing: An Integrated Model of Scripts, Lexicon, and Memory*. MIT Press, Cambridge, MA.

Mitra, S., and Pal, S.K. (1994): Self-organizing neural network as a fuzzy classifier. In *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 24, no. 3, pp. 385-398.

Ritter, H., and Kohonen, T. (1989): Self-organizing semantic maps. *Biological Cybernetics*, vol. 61, no. 4, pp. 241-254.

Ritter, H., and Kohonen, T. (1990): Learning 'semantotopic maps' from context. In *Proceedings of IJCNN'90, International Joint Conference on Neural Networks*, vol.I, pp. 23-26. Lawrence Erlbaum, Hillsdale, NJ.

Scholtes, J.C. (1991a): Kohonen feature maps in full-text data bases: A case study of the 1987 Pravda. In *Proceedings of. Informatiewetenschap 1991*, Nijmegen, pp. 203-220, STINFON, Nijmegen, Netherlands, 1991.

Scholtes, J.C. (1991b): Unsupervised learning and the information retrieval problem. In *Proceedings of IJCNN'91, International Joint Conference on Neural Networks*, pp. 18-21. IEEE Service Center, Piscataway, NJ, 1991.

Scholtes, J.C. (1993): *Neural Networks in Natural Language Processing and Information Retrieval*. PhD thesis, Universiteit van Amsterdam, Netherlands.

Schütze, H. (1992): Dimensions of meaning. In *Proceedings of Supercomputing '92*, pp. 787-796, Minneapolis, MN.

Streeter, L.A., and Lochbaum, K.E. (1988): Who knows: a system based on automatic representation of semantic structure. *Proceedings of RIAO'88*, Cambridge, MA.

Ultsch, A. (1993): Self-organizing neural networks for visualization and classification. In O.Opitz, B. Lausen and R. Klar (eds.) *Information and Classification*, pp. 307-313, Springer-Verlag, Berlin.