In Sincak P. et al (Eds.), Cognition and Artificial Life III, Stara Lesna, Slovakia, pp. 13-18, 2003

What is the structure of human lexical system?

Igor Farkaš [1]
Department of Psychology
University of Richmond, USA

Abstract: This paper pinpoints some current issues related to computational approaches to semantic lexical representations and language structure. First, we discuss context-based methods that allow to generate word meanings based on a window sliding through big text corpora. Word associations as predicted by these vector-space models are then assessed with regard to results from psychological priming experiments that provide solid evidence on how humans associate words with one another. This comparison is then extended to the lexical system projected topographically into lexical maps. Finally, we provide evidence for the possible lexical structure as suggested by studies of language networks that possess small world and scale-free properties. Since these language network models suggest a much richer word connectivity structure than that of resulting from vector-space models, potentially including spurious ssociations, we suggest that the connectivity pattern of the human lexical semantic system might lie in between the two bounds.

Representing lexical semantics

On the formal level of description, the human lexical system may be thought of as a graph whose nodes stand for words and whose links represent associations between words. Within this framework, two main issues emerged as foci of the research: (1) how to represent the lexical (word) semantics, and (2) how the words (concepts) are associated with one another. Both issues can be tackled with computational models, but in both cases, the model predictions should be confronted with empirical evidence about human linguistic behavior to assess psychological and cognitive plausibility of these models.

How to represent semantics computationally has been a difficult problem for many years, and as yet, there is no consensus as to exactly what is stored and how. With the rise of cognitive modeling, several approaches to representing lexical semantics have been recently introduced. In their nature, they all essentially are vector space models, but they differ in how the word vectors are created and what their components reflect [1]. For instance, one can take advantage of explicit human judgements and use word features consistently generated by most subjects. Alternatively, researchers themselves can generate these semantic vectors using their intuition. As a third method, random vectors are used, with an argument that for modeling some processes (e.g., arbitrary mapping betweeen semantics and phonology), the actual components of the semantic vectors do not matter. Last, an automatic method usable in large scale tasks was introduced that uses WordNet [2], a lexigraphic database, for creating semantic vectors of words in question [3].

All the above mentioned methods (except random vectors) generate features that point to characteristics of the corresponding word. Hence, the representations are grounded in the real world. However, recent work in computational linguistics and cognitive psychology (see e.g., [4] and [5], respectively, and references therein) has shown that interesting features of words (including semantic) can also be captured merely from word

[1] On leave from Insitute of Measurement Science, Slovak Academy of Sciences, Bratislava, Slovakia

co-occurrence statistics. Various modifications of this approach have been proposed, as for what is considered a word context. For instance, in Latent Semantic Analysis (LSA, [5]), word occurrence across various documents (contexts) is counted, whereas in Hyperspace Analogue to Language (HAL, [6]), the context is represented by a fixed window sliding through the corpus. Manipulating window parameters (such as size, shape) and distance metrics affects the representations. For example, it has been observed [1] that syntactic information is better captured with a very short window, whereas semantic features may require larger windows. It has also been observed that left word context tends to provide more information then the right context.

In my recent work on modeling early lexical aquisition [7], I also derived context-based word representations from the parental CHILDES corpus [8]. The simulations supported the latter of the above two claims (left context is better than the right), but not the former: increasing the window radius did not improve the compactness of all grammatical/semantic categories. In addition, I combined word context information with Wordnet-based semantic features of words [3] by concatenating both vectors for each word. This led to two improvements: (1) overall accuracy of word representations (as measured by clustering properties of word categories) was increased and (2) semantic represenions became grounded in real world (thus circumventing a common objective against using purely word co-occurrence methods as being cognitively implausible).

To make word context methods computationally effective, various methods were advocated to reduce the number of dimensions of co-occurrence vectors, while approximately preserving the data structure. For instance, in LSA [5] 300 dimensions are extracted using singular-value decomposition of the word-document cooccurrence matrix. In HAL [6], 200 vector components with highest variability are selected. In our model of growing semantic map for lexical acquisition [9] we linearly project the high-dimensional word vectors (whose dimensionality grows linearly with lexicon size) down to 100-dimensions using random mapping procedure [10].

Account for semantic priming

Whatever the dimensionality of word representations, semantically similar words are expected to be represented by two points being close in vector space. Vector space models have proved very successful models of semantic memory. The primary test of these models is their accountibility for priming effects. Semantic priming paradigm [11], as one of the most widely used in psychological studies of memory and language, demonstrates how the recognition of the target word (measured by the reaction time) can be facilitated when preceded by a related (priming) word. In other words, the reaction time taken to recognize the target word is shorter when the target is preceded by related than when preceded by unrelated word. Priming has been observed between word pairs that are from the same semantic category (e.g., *run-jump*) or are functionally related (e.g., *hammer-nail*) [12]. In testing whether vector space models can account for priming effects, spatial proximity (cosine of two vectors) of related prime word and target is compared to that of unrelated prime words. If the related prime word is closer to the target than unrelated prime word, this is taken as account for priming effect. Vector space models demonstrated the priming effect account for various types of semantic relatedness between words [5, 13, 14].

Semantic memory maps

When one focuses on how such a semantic vector space model could be mapped onto a human cortex, the question arises as for how well the structure can be preserved when compressed in two dimensions. There exists experimental evidence for topographic organization of information in the brain as a potentially universal representation mechanism used in various sensory modalities [15]. It has also been suggested [10] that there may exist semantic topographic maps in the brain in which words would be ordered according to their semantic relationships. This hypothesis is also supported by a numerous neuropsychological evidence on category-specific naming deficits observed in patients who cannot name objects belonging to a certain semantic category (but can for other categories) as a consequence of a stroke or other brain disorder (e.g., [16]).

In trying to validate such semantic maps with respect to psychological perspective, Lowe [17] has shown that priming effects both between nouns from the same semantic category and for funcionally related nouns can be replicated when looking at word distances in the 2D map (extracted by Generative Topographic Mapping [18]), rather than in original high-dimensional space. In our experience with CHILDES data using 500 words, most of the semantic categories also remained mostly preserved in the self-organizing map [7]. However, what remains roughly preserved are mostly semantic associations beween words belonging to the same category, whereas other (e.g., functional) associations may be lost. Most probably, more links need to be added to the structure in the form of lateral, long-range connections. Also, the literature on priming effects focuses on relations between nouns (paradigmatic associations), but the syntagmatic relations between words can also lead to priming effects (e.g., *dog-bark*) and such effects could hardly be accounted for in the semantic map without lateral connections. In addition, it is more likely that the lexicon could be organized as a set of interconnected and hierachical maps rather than one huge map. Neuropsychological evidence suggests that e.g., grammatical categories such as nouns and verbs are processed by separate brain regions [19] and so may be the superordinate and subordinate categories of nouns [20].

Word form and word meaning

There exists solid psychological and neurobiological support for the theory that word forms (ortographic and phonological) and word meanings could be subserved by separate neural systems [21]. To account for existence of homographs (ortographically ambiguous words) and synonyms, there has to exist a many-to-many mapping between ortographic word form and meaning. As a consequence, a complication arises for context space models because what they actually derive for ambiguous words is a blended meaning averaged over different contexts in which the word is found in the corpus (the method cannot differentiate between unrelated contexts and collect separate statistics for various meanings of polysemous words). For some word forms, the blended meaning could be referred to as a *root* meaning, (e.g., 42 meanings of the verb *run* according to WordNet), but in case of other word forms the blended meaning has no interpretation (e.g., *bank* as a financial institution, or a river bank).

Small world of human language

Unlike vector space methods in which geometrical structure (Euclidean distance or cosine between two vectors) determines the similarities between words considered as concepts, a different connectivity pattern emerges when one treats the lexicon as a graph whose connectivity results from coocurrences of word forms. Specifically, networks of human language became recently studied from the perspective of complex systems, in which the focus is on the graph structure. Hence, the actual coordinates of nodes (word meanings) are irelevant. There have been a few models proposed. One method connects words based on their cooccurrence across sentences within a window of radius two [22]. In another approach [23], using Thesaurus electronic dictionary, two nodes are connected if one of the corresponding words is the listed in the entry of the other one. Despite differences in resulting connectivities, both studies lead to the same conclusions. A language network, like many other complex systems (e.g., world-wide web, telecommunication network or citation networks of scientists), display two important characteristics: the so called *small-world* property and *scale-free* structure.

Small-world property [24] is characterized by the combination of highly clustered neighborhoods and a short average path-length between two nodes. This means that any two words can be connected via only a few intermediate words and this is possible thank to a small subset of polysemous words that have many neighbors and serve as hubs within the network. Scale-free structure [25], on the other hand, is defined by an algebraic behavior in the probability distribution $P(k)$ of k, the number of links at a node in the network. In a scale-free network, the distribution of node connectivities follows a power function, rather than an exponential function.

Both properties shed light on appealing attributes of language. Specifically, small-world phenomenon allows to appreciate the efficient retrieval within an associative network: on the one hand, similar words are clustered together, but even very different words are never separated by more than a few links, which guarantess fast search [23]. The scale-free property suggests interesting mechanisms that may have shaped the evolution of the word network [26].

As the other side of the coin, it can be seen that connectivity pattern resulting from these language network results in much richer network of association than is possible with vector-space model in which connections are based on Euclidean distances (points whose distances is below predefined threshold become connected to yield a graph). As a matter of fact, some researchers [27] questioned the appropriateness of vector (Euclidean) spaces for conceptual networks because they impose un upper bound on the number of points that can share the same nearest neighbor (which applies to map representations, too). On the other hand, language networks with small-world property may assume word connections that are truly spurious, and most probably, would not show priming effect. For instance, in case of Cancho and Solé's model [22], it is unlikely that all neighboring word pairs in the text would be neighbors in the mental lexicon as well. Similar doubts could be raised when considering Motter et al's model [23].

Word's neighbors

To computationally assess the above considerations, we calculated the average number of word neighbors (k) for our vector-space model [9] and for a small-world language network of Cancho and Solé [22], both of which are based on a sliding window that determines the network structure (we considered window radius 1 and 2). We took 2473 most frequent words (corresponding to token frequency of 16 and higher per million words) from parental CHILDES database containing approximately 3.1 million words. Whereas the connectivity in Cancho and Solé's model directly results from the graph, in vector-

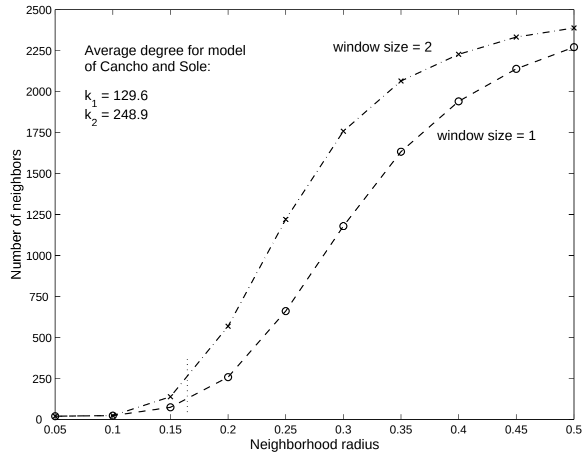

Figure 1: Average number of word neighbors as a function of neighborhood radius (relative to Euclidean distance between two most distant words) evaluated for our vector-space model on CHILDES corpus (2473 most frequent words). As a reference, the average number of nearest neighbors (degree) obtained for the same corpus based on the language network model of Cancho and Solé is shown.

space model k depends on the considered radius n_r of the hypersphere surrounding the target word. This radius was evaluated as a fraction of maximum Euclidean distance between two words in the lexicon (hence, $n_r = 0.5$ covers the whole space if the word is in the center of word space). As shown in Figure 1, this function is nonlinear, and as could be expected, k grows faster in mid-range for $rad = 2$. Corresponding small-world networks respectively yield $k_1 = 129.6$ and $k_2 = 248.9$. To match these averages, the diameter of the neighborhood hypersphere would have to be one third of that of the word space (for $n_r = 0.17$; denoted by a short dotted line), to allow for the same k as the small-world model. Since vector-space models connect word meanings whereas small-world models connect word forms, it is difficult for the former to account for higher number of nearest neighbors in case of polysemous words without allowing larger n_r. Since considered n_r has to be the same for all words, as a trade-off, larger n_r increases the likelihood of including spurious neighbors.

Therefore, we conclude by suggesting that vectors-space models may underestimate the actual neighborhood structure, whereas small-world models appear to overestimate it. Therefore, it is possible that the connectivity pattern of the human lexical system will lie somewhere between the two bounds.

References

[1] J.P. Levy and J.A. Bullinaria. Explorations in the derivation of word co-occurrence statistics. *South Pacific Journal of Psychology*, 10(1):99–111, 1998.

[2] C. Fellbaum. *WordNet: An Electronic Lexical Database.* MIT Press, Cambridge, MA, 1998.

[3] M.W. Harm. Building large scale distributed semantic feature sets with WordNet. Technical Report PDP.CNS.02.1, Carnegie Mellon University, Pittsburg, PA, 2002.

[4] I. Dagan, L. Lee, and F. Pereira. Similarity-based models of word cooccurrence probabilities. *Machine Learning*, 34(1-3):43–69, 1999.

[5] T.K. Landauer and S.T. Dumais. A solution to Plato's problem: the latent semantic analysis theory of induction and representation of knowledge. *Psychological Review*, 104:211–240, 1997.

[6] C. Burgess and K. Lund. Modelling parsing constraints with high-dimensional semantic space. *Language and Cognitive Processes*, 12(2/3):177–210, 1997.

[7] I. Farkaš. Lexical acquisition and developing semantic map. Submitted.

[8] P. Li, C. Burgess, and K. Lund. The acquisition of word meaning through global lexical co-occurrences. In E. Clark, editor, *Proceedings of the 30th Child Language Research Forum*, pages 167–178. Stanford, CA: Center for the Study of Language and Information, 2000.

[9] I. Farkaš and P. Li. Modeling the development of lexicon with a growing self-organizing map. In H.J. Caulfield et al, editor, *Proceedings of the 6th Joint Conference on Information Sciences*, pages 553–556, Research Triangle Park, NC, 2002.

[10] H. Ritter and T. Kohonen. Self-organizing semantic maps. *Biological Cybernetics*, 61:241–254, 1989.

[11] D.E. Meyer and R.W. Schvaneveldt. Facilitation in recognising pairs of words: Evidence of a dependence between retrieval oparations. *Journal of Experimental Psychology*, 90:227–235, 1971.

[12] H.E. Moss, L.K. Tyler R.K. Ostrin, and W.D. Marslen-Wilson. Accessing different types of lexical semantic information: Evidence from priming. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 21(4):863–883, 1995.

[13] K. Lund, C. Burgess, and R.A. Atchley. Semantic and associative priming in high-dimensional semantic space. In *Proceedings of the 17th Annual Conference of the Cognitive Science Society*, pages 660–665, 1995.

[14] S. McDonald and W. Lowe. Modelling functional priming and the associative boost. In *Proceedings of 20th Annual Conference of the Cognitive Science Society*, pages 675–680. LEA, 1998.

[15] E.I. Knudsen, S. du Lac, and S.D. Esterly. Computational maps in the brain. *Annual Review of Neuroscience*, 10:41–65, 1987.

[16] A. Caramazza, A. Hillis, E.C. Leek, and M. Miozzo. The organization of lexical knowledge in the brain: Evidence from category- and modality-specific deficits. In L.A. Hirschfeld and S.A. Gelman, editors, *Mapping the Mind: Domain Specificity in Cognition and Culture*, pages 68–84. Cambridge University Press, Cambridge, 1994.

[17] W. Lowe. What is the dimensionality of human semantic space? In R.M. French and J.P. Sougné, editors, *Proceedings of the 6th Neural Computation and Psychology Workshop*, pages 303–311, London, 2000. Springer Verlag.

[18] C.M. Bishop, M. Svensén, and C.K.I. Williams. GTM: The generative topographic mapping. *Neural Computation*, 10(1):215–235, 1998.

[19] A.E. Hillis and A. Caramazza. Representation of grammatical categories of words in the brain. *Journal of Cognitive Neuroscience*, 7(3):396–407, 1995.

[20] A. Caramazza, A.E. Hillis, B.C. Rapp, and C. Romani. The multiple semantics hypothesis: Multiple confusions? *Cognitive Neuropsychology*, 7:161–190, 1990.

[21] R.C. Martin. Language processing: Functional organization and neuroanatomical basis. *Annual Review of Psychology*, 54:55–89, 2003.

[22] R.F. i Cancho and R.V. Solé. The small world of human language. *Proceedings of the Royal Society of London: Series B*, 268:2261–2265, 2001.

[23] A.E. Motter, A.P.S. de Moura, Y.-C. Lai, and P. Dasgupta. Topology of the conceptual network of language. *Physical Review E*, 65:065102, 2002.

[24] D.J. Watts and S.H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393:440–442, 1998.

[25] A.-L. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286:509–511, 1999.

[26] S.N. Dorogovtsev and J.F.F. Mendes. Language as an evolving word web. *Proceeedings of the Royal Society of London: Series B*, 268:2603–2606, 2001.

[27] A. Tversky and J.W. Hutchinson. Nearest neighbor analysis of psychological spaces. *Psychological Review*, 93(1):3–22, 1986.