Very Large Neural Networks for Word Sense Disambiguation

Nancy M. IDE (*) and **Jean VERONIS** (* and **)

*Department of Computer Science
VASSAR COLLEGE
Poughkeepsie, New York 12601 (U.S.A.)

** Groupe Représentation et Traitement des Connaissances
CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE
31, Ch. Joseph Aiguier
13402 Marseille Cedex 09 (France)

European Conference on Arificial Intelligence, ECAI'90, Stockholm, August 1990

Abstract

The use of neural networks for word sense disambiguation has been suggested, but previous approaches did not provide any practical means to build the proposed networks. Therefore, it has not been clear that the proposed models scale up to realistic dimensions. We demonstrate how textual sources such as machine readable dictionaries can be used to automatically create Very Large Neural Networks which embody a simple local model and contain several thousands of nodes. Our results show that such networks are able to disambiguate words at a rate three times higher than chance, and thus provide real-size validation for connectionnist approaches to word sense disambiguation.

1. Introduction

Most words in a language have several different senses, and the comprehension of language demands that the exact sense of each word in a given context be determined. The use of neural networks to disambiguate word senses has been proposed by Cottrell and Small (1983) and Waltz and Pollack (1985). However, these networks are created manually and are therefore necessarily limited in size (no more than a few dozen nodes). It is not clear that these models will scale up to realistic dimensions. In this paper we describe means to automatically create Very Large Neural Networks, containing several thousands of nodes, from machine-readable dictionaries. This technique enables real size experiments with neural networks for word sense disambiguation, which in turn provide insight into connectionist models for natural language comprehension.

2. The Model

We build VLNNs utilizing definitions in the *Collins English Dictionary*. The connections in the network reflect the semantic relation between a word and the words used to define it. All of the knowledge represented in the network is automatically generated from a machine-readable dictionary, and therefore no hand coding is required. The lexicon and the knowledge it contains potentially cover all of English (the *Collins English Dictionary* contains 90,000 entries), and, as a result, this information can potentially be used to help disambiguate unrestricted text.

2.1. Network Topology

We are using a *local* model, that is, each entity (word or sense) is represented by a node in the network. Each lexical entry is represented by a complex grouping of nodes, consisting of a central node (or *word node*) connected by activatory links to *sense nodes* that represent the different senses (definitions) of this word in the *Collins English Dictionary*. The different sense nodes for a given word are completely interconnected by inhibitory links. For each connection from a node *i* to a node *j*, there is a reciprocal connection from node *j* to node *i*. Each sense node is connected by activatory links to word nodes representing the words that appear in its definition, and these words are connected to their sense nodes in turn, etc. This process is repeated a number of times, creating an increasingly complex and interconnected network (figure 1).

Ideally, the network would include the entire dictionary, but for practical reasons we limit the number of repetitions to 2 and thus restrict the size of the network to a few thousand nodes and 10 to 20 thousand transitions. All words in the network are reduced to their lemmas, and grammatical words are excluded.

2.2. Dynamic Behavior of the Network

When the network is run, the input word nodes are activated first. Then each input word node sends activation to its sense nodes, which in turn send activation to the word nodes to which they are connected, and so on throughout the network for a number of cycles. At each cycle, word and sense nodes also receive *feedback* from connected nodes. Sense nodes for the same word, which are in competition, inhibit one another. Feedback and inhibition cooperate in a "winner-take-all" strategy to activate increasingly related word and sense nodes and deactivate the unrelated or weakly related nodes. Eventually, after a few dozen cycles, the

The authors would like to acknowledge the contributions of Stéphane Harié, Gavin Huntley, and Michael Scharf to the work presented in this paper.

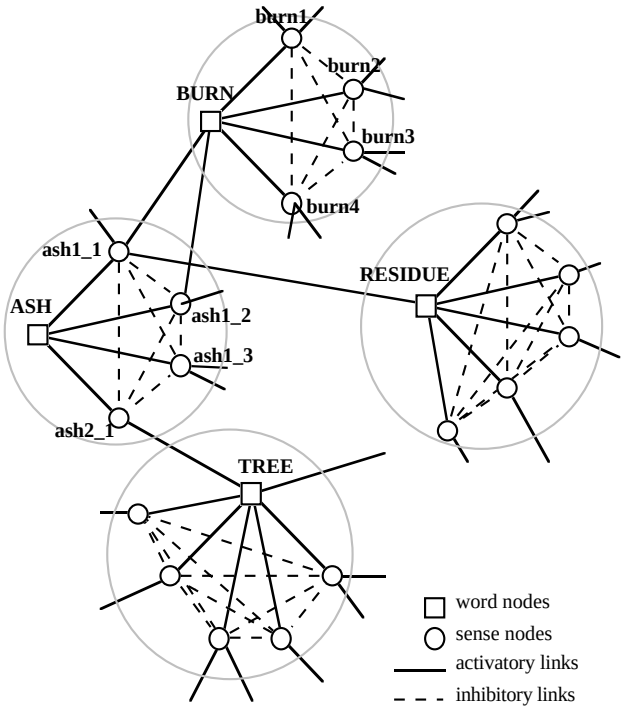

Figure 1. Topology of the network

network stabilizes in a configuration where only the sense nodes with the strongest relations to other nodes in the network are activated. Because of the "winner-take-all" strategy, at most one sense node per word will ultimately be activated.

The activation model is derived from McClelland and Rumelhart (1981). At any instant *t*, a node *i* has an activation level $a_i(t)$ with a real value. If a node has a positive activation level, it is said to be *active*. If the node receives no input from its neighbors, it tends to return to a *resting level* r_i, more or less quickly depending on decay rate θ_i. The active neighbors of a node affect its activation level either by excitation or inhibition, depending on the nature of the link. Each connection from a node *j* towards another node *i* has weight w_{ji}, which has a positive value for excitatory links and a negative value for inhibitory links (w_{ji} can be different from w_{ij}). The total input $n_i(t)$ of input from its neighbors on node *i* is the inner product of the vector of the neighbors' activation levels by the connection weight vector:

$$n_i(t) = \sum_j w_{ji} a_j(t)$$

This input is constrained by a "squashing function" which prevents the activation level from going beyond a defined minimum and maximum. A node becomes increasingly difficult to activate as its level approaches the maximum, and increasingly difficult to inhibit as it approaches the minimum. In mathematical terms, when the overall input is *activatory*, that is, when $n_i(t) > 0$, the actual effect on node *i* is determined by:

$$\varepsilon_i(t) = n_i(t)\ (\max - a_i(t)).$$

When the overall input is *inhibitory*, the actual effect is determined by:

$$\varepsilon_i(t) = n_i(t)\ (a_i(t) - \min).$$

The new activation level of a node *i* at time $t + \Delta t$ corresponds to the activation of this node at time *t*, modified by its neighbors' effect and diminished by the decay $\theta_i(a_i(t) - r_i)$:

$$a_i(t + \Delta t) = a_i(t) + \varepsilon_i(t) - \theta_i(a_i(t) - r_i).$$

Finally, each node is affected by a threshold value τ_i, below which it remains inactive.

Connection weights are fixed *a priori*. In early experiments, they were the same for all connections, but we discovered that "gang effects" appear due to extreme imbalance among words having few senses and hence few connections, as well as words containing up to 80 senses and several hundred connections, and that therefore dampening is required. In our current experiments, connection weights are determined by a simple decreasing function of the number of outgoing connections from a given node.

3. Testing the Network

3.1. Method

To test the network, we used a simplified case in which the input consists of one ambiguous word and a second "context word". Using only two words as input enabled us to better understand the behavior of the network, although we must bear in mind that with only one word of context, the disambiguation task is more difficult. Our more recent experiments show that additional context improves the results significantly.

We tested the network with 23 clearly ambiguous words in different contexts. For each of these words, at least two homographs (with unrelated etymologies--for example, ash_1 as residue and ash_2 as tree) exist in the *Collins English Dictionary*. In turn, each homograph may have several different senses (ash_2, for instance, has a sense defining the ash tree itself and another defining the wood of that tree). On average, each word in our list has 2.9 homographs and 6.7 senses.

WORD	CONTEXT 1	CONTEXT 2	CONTEXT 3
ash_1	residue	fire	tobacco
ash_2	tree	branch	lawn
bay_1	peninsula	cape	sail
bay_2	wall	window	house

Figure 2. A sample of the corpus

For each of the 23 words, we constructed 2 groups of 3 context words, using examples from the *Brown Corpus of American English* whenever possible. Each of the two groups is semantically related to one homograph (figure 2). The three context words are differentiated on the basis of an increasing minimum pathlength to the ambiguous word within the network. Thus the final experimental corpus consists of 138 word pairs.

3.2. Results

The network was run on the 138 word pairs, and the number of correctly disambiguated senses was computed. In a few cases, two different senses for the same homograph were accepted as correct, as with *sage/rosemary*, where *sage* can be regarded as a plant or the spice. In 71.74% of the cases, the correct sense was indentified by the network, which is 3 times better than chance (23.63%).

On a few occasions, the network identified the correct homograph, but identified the wrong sense for that homograph. For instance, *reel/camera* activated the correct homograph $reel_1$ (cylindrical object) as opposed to $reel_2$ (Scottish dance), although sense 2 of $reel_1$ was selected, which is a restricted sense pertaining to angling. Therefore, the network's ability to determine the correct homograph was slightly better than for senses, at 73.19%.

3.3. Discussion

Although our model is only preliminary, the network successfully disambiguates word senses at a rate 3 times better than random. This is especially encouraging when we consider that in some of our sample input pairs, the context word is itself ambiguous (for instance, *bay/cape*), and the network correctly disambiguates both words.

There may be several factors contributing to those cases where the network failed to identify the correct sense. First, the input provides only one context word, which, as mentioned earlier, makes the task particularly difficult. In addition, the parameters used in these experiments are a first approximation, and it remains to be seen what the effect of changing parameters such as the amount of feedback or inhibition, etc. could have on the results. Most experiments in connectionism show that neural networks are extremely sensitive to tuning of this kind. Also, our experiments were run using a dictionary whose sense differentiations are extremely precise, often specifying subtle distinctions among senses and identifying very rare or obsolete senses. The network's behavior may improve with a dictionary providing more clear-cut information, such as a learner's dictionary, which would be more than adequate for most language processing tasks.

In our current model, all sense nodes for all homographs of a given word are attached directly to its word node, and therefore they are all mutually inhibited. Instead, senses within the same homograph (which are semantically related) should inhibit each other less than senses between different homographs (which are not semantically related). Therefore, we may see some improvement in the results if the model is modified to add intermediary nodes accounting for different homographs. The same problem exists when senses are split into subsenses, which the current model could also be modified to take into account, leading to a more complex, hierarchical unit for each word.

Many of the disambiguation failures in our corpus result from remote relations between the two input words. As noted above, we iterate the process of building word-to-sense-to-word links 2 times in our current implementation. The results may be improved with additional iterations or, ideally, with a single network covering the entire dictionary. Our experiments show that with appropriate activation threshold values, only very small sections of this ideal network would be activated at any given time. We are currently working on a large scale implementation that would keep only those parts of the network which are activated in main memory.

In some cases, however, the appropriate links between words are not found in the dictionary at all. Because the purpose of dictionaries is to define individual lexical entries, much broad contextual or world knowledge is not represented in dictionary definition texts. For instance, it is interesting to note that there is no semantic path between *lawn* and *house* in the *Collins English Dictionary*, although it is clear that the connection is part of human experience. One possible solution to this problem is to combine informaton from several dictionaries. Beyond this, we may use collocational information from text corpora such as the *Brown Corpus*, in which, for example, *house* co-occurs with *lawn* in 3 of the 14 occurrences of the latter.

4. Conclusion

Although our model is very preliminary, we have seen interesting real-scale results which provide some practical validation for suggested connectionist models applied to language processing tasks. Previous approaches did not provide any practical means to build the proposed networks. We demonstrate how textual sources such as machine readable dictionaries can be used, without sophisticated pre-processing, to build Very Large Neural Networks. Our results provide in turn some positive evidence in the debate over whether the enormous body of lexical knowledge encoded in dictionaries can be exploited for natural language processing.

References

COTTRELL, G. W., SMALL, S. L. (1983). A connectionist scheme for modelling word sense disambiguation. *Cognition and Brain Theory*, 6, 89-120.

MCCLELLAND, J. L., RUMELHART, D. E. (1981). An interactive activation model of context effects in letter perception: Part I. An account of basic findings. *Psychological Review*, 88, 375-407.

WALTZ, D. L., POLLACK, J. B. (1985). Massively Parallel Parsing: A Strongly Interactive Model of Natural Language Interpretation. *Cognitive Science*, 9, 51-74.