LARGE NEURAL NETWORKS FOR THE RESOLUTION OF LEXICAL AMBIGUITY

JEAN VÉRONIS, NANCY IDE

Groupe Représentation et Traitement des Connaissances,
Centre National de la Recherche Scientifique,
31, Chemin Joseph Aiguier, 13402 Marseille Cedex 09, France

and

Department of Computer Science, Vassar College
Poughkeepsie, New York 12601, U.S.A.

1. Introduction

Many words have two or more very distinct meanings. For example, the word *pen* can refer to a writing implement or to an enclosure. Many natural language applications, including information retrieval, content analysis and automatic abstracting, and machine translation, require the resolution of lexical ambiguity for words in an input text, or are significantly improved when this can be accomplished.[1] That is, the preferred input to these applications is a text in which each word is "tagged" with the sense of that word intended in the particular context. However, at present there is no reliable way to automatically identify the correct sense of a word in running text. This task, called *word sense disambiguation,* is especially difficult for texts from unconstrained domains because the number of ambiguous words is potentially very large. The magnitude of the problem can be reduced by considering only very gross sense distinctions (e.g., between the pen-as-implement and pen-as-enclosure senses of *pen,* rather than between finer sense distinctions within, say, the category of pen-as-enclosure--i.e., enclosure for animals, enclosure for submarines, etc.), which is sufficient for many applications. But even so, substantial ambiguity remains: for example, even the relatively small lexicon (20,000 entries) of the TRUMP system, which includes only gross sense distinctions, finds an average of about four senses for each word in sentences from the *Wall Street*

Journal (McRoy, 1992). The resulting combinatoric explosion demonstrates the magnitude of the lexical ambiguity problem.

Several different kinds of information can contribute to the resolution of lexical ambiguity. These include word-specific information such as morphological information, part of speech (the syntactic category of the word), relative sense frequency (preferred sense, either generally of for the determined domain), semantic features (the semantic components, often drawn from a potentially large set of primitives, that contribute to meaning); as well as contextual information such as syntactic role (e.g., a particular sense may be the only one allowed as the object of a given preposition), role-related preferences (selectional restrictions defining relations between a noun and verb), semantic relations (most usually, senses of or associations with surrounding words), etc. It has recently been suggested that an effective word sense disambiguation procedure will require information of most or all these types (McRoy, 1992). However, most methods utilize only one or two of the potential information sources listed above. One common approach follows in the case-based tradition of Fillmore (1968) and Schank (1975), and involves matching role restriction specifications with semantic features in order to choose the appropriate senses for other words in the sentence (for example, the AGENT role for a given verb must have the feature ANIMATE) (see, for example, Wilks and Fass, 1990; Fass, 1988). Another approach to word sense disambiguation uses semantic context as an indicator of sense, and involves the traversal of semantic networks that indicate word and/or concept associations, typically utilizing marker-passing techniques (Charniak, 1983; Hirst, 1987; Norvig, 1989). A similar approach implements connectionist models of human meaning representation based on semantic features, using a spreading activation strategy (Cottrell and Small, 1983; Waltz and Pollack, 1985).

All of these techniques require the prior creation of a potentially complex lexical and/or semantic knowledge base, which poses theoretical and practical problems. Designing role-related systems and identifying appropriate feature sets for unrestricted language is not straightforward in the state of the art. In any case, it is often impractical to manually encode this information for a lexicon of substantial size[2]. As a result, several studies have attempted to use ready-made information sources such as machine-readable versions of everyday dictionaries (Lesk, 1986; Wilks *et al.*, 1990) and word-aligned bi-lingual corpora (Gale, Church, and Yarowsky, to appear) for word sense disambiguation, in order to avoid the high costs of creating large lexical and semantic knowledge bases.

We describe here a method for word sense disambiguation that builds upon both connectionnist models and the work of Lesk and Wilks. We automatically construct very large neural networks from definition texts in machine-readable dictionaries, and therefore no hand coding of large-scale resources is required. Further, the knowledge contained in the network therefore potentially covers all of English (90,000 words), and as a result, this information can potentially be used to help disambiguate unrestricted text. Creation of the networks requires only minimal pre-processing of definition texts, involving the elimination of non-content words and lemmatization. There are nodes in the network for the root forms of all words appearing in the dictionary, linked to concept nodes corresponding to their senses. Concept nodes are in turn linked to the root forms of words appearing in their definitions. The fundamental assumption underlying the semantic knowledge represented in these networks is that there are significant semantic relations between a word and the words used to define it. The connections in the network reflect these relations. There is no indication within the network of the *nature* of the relationships, although the presence of words with important and relatively fixed semantic relations to their headwords in dictionary definitions is well-known, and much work has been applied to identifying and extracting this information (see, for instance, Amsler, 1980; Calzolari, 1984; Chodorow, Byrd and Heidorn, 1985; Markowitz, Ahlswede and Evens, 1986; Byrd *et al.,* 1987; Véronis and Ide, 1991; Nakamura and Nagao, 1988; Klavans, Chodorow, and Wacholder, 1990; Wilks *et al.,* 1990). Such information is not systematic or even complete, and its extraction from machine-readable dictionaries is not always straightforward. However, it has been shown that in its base form, information from machine-readable dictionaries can be used, for example, to find subject domains in texts (Walker and Amsler, 1986). More importantly, it is not clear that knowing the nature of the relationships would significantly enhance the word sense disambiguation process.

Input to the network consists of words from running text. There is no requirement for complex processing of the input text, which means that the method can be applied to unrestricted text. This is in contrast to the role-related approach, which demands the a priori identification of syntactic elements, a costly process which in the current state of the art is almost impossible for unrestricted text. We apply a spreading activation strategy which ultimately identifies the sense of each input word that shares the most relations with other input (context) words. As such, our work, like the semantic network and connectionnist approaches mentioned above, relies on semantic context for disambiguation, and thus follows from work in lexical cohesion which shows that in order to create text unity, people repeat words in a given conceptual category or which are semantically associated (Halliday and Hasen, 1976; Morris and Hirst, 1991). Output from the network is the text with each word tagged with its appropriate sense (as identified in the dictionary used to construct the network), which can in turn serve as input to other processes.

This work has been carried out in the context of a joint project of Vassar College and the Groupe Représentation et Traitement des Connaissances of the Centre National de la Recherche Scientifique (CNRS), which is concerned with the construction and exploitation of a large lexical data base of English and French.

2. Previous work

2.1. Machine-readable dictionaries for disambiguation

The most general and well-known attempt to utilize information in machine-readable dictionaries for word sense disambiguation is that of Lesk (1986), which computes the degree of overlap--that is, number of shared words--in definition texts of words that appear in a ten-word window of context. The sense of a word with the greatest number of overlaps with senses of other words in the window is chosen as the correct one. For example, consider the definitions of *pen* and *sheep* from the *Collins English Dictionary* (*CED*°, the dictionary used in our experiments, in figure 1.

If these two words appear together in context, the appropriate senses of *pen* (2.1: *enclosure*) and *sheep* (1: *mammal*) will be chosen because the definitions of these two senses have the word *domestic* in common. However, with one word as a basis, the relation is tenuous and wholly dependent upon a particular dictionary's wording. The method also fails to take into account less immediate relationships between words. As a result, it will not determine the correct sense of *pen* in the context of *goat.* The correct sense of *pen* (2.1: *enclosure*) and the correct sense of *goat* (1: *mammal*) do not share any words in common in their definitions in the *Collins English Dictionary;* however, a strategy which takes into account a longer path through definitions will find that *animal* is in the definition of *pen* 2.1, each of *mammal* and *animal* appear in the definition of the other, and *mammal* is in the definition of *goat* 1.

Similarly, Lesk's method would also be unable to determine the correct sense of *pen* (1.1: *writing utensil*) in the context of *page,* because seven of the thirteen senses of *pen* have the same number of overlaps with senses of *page.* Six of the senses of *pen* share only the word *write* with the correct sense of *page* (1.1: *leaf of a book*). However, *pen* 1.1 also contains words such as *draw* and *ink*, and *page* 1.1 contains *book*, *newspaper, letter,* and *print.* These other words are heavily interconnected in a complex network which cannot be discovered by simply counting overlaps. Wilks *et al.* (forthcoming) build on Lesk's method by computing the degree of overlap for related word-sets constructed using co-occurrence data from definition texts, but their method suffers from the same problems, in addition to combinatorial problems that prevent disambiguating more than one word at a time.

pen[1] *n.* **1.** an implement for writing or drawing using ink, formerly consisting of a sharpened and split quill, and now of a metal nib attached to a holder. **2.** the writing end of such an implement; nib. **3.** style of writing. **4. the pen. a.** writing as an occupation. **b.** the written word. **5.** the long horny internal shell of a squid. ~*vb.* **6.** *(tr.)* to write or compose.
pen[2] *n.* **1.** an enclosure in which domestic animals are kept. **2.** any place of confinement. **3.** a dock for servicing submarines. ~*vb.* **4.** *(tr.)* to enclose or keep in a pen.
pen[3] *n.* short for **penitentiary**.

pen[4] *n.* a female swan.

sheep *n.* **1.** any of various bovid mammals of the genus *Ovis* and related genera, esp. *O. aries* (**domestic sheep**) having transversely ribbed horns and a narrow face. There are many breeds of domestic sheep, raised for their wool and for meat. **2. Barbary sheep.** another name for **aoudad. 3.** a meek or timid person, esp. one without initiative. **4. separate the sheep from the goats.** to pick out the members of any group who are superior in some respects.

goat *n.* **1.** any sure-footed agile bovid mammal of the genus *Capra*, naturally inhabiting rough stony ground in Europe, Asia, and N Africa, typically having a brown-grey colouring and a beard. Domesticated varieties (*C. hircus*) are reared for milk, meat, and wool. **2.** short for **Rocky Mountain goat 3.** *Informal.* a lecherous man. **4.** a bad or inferior member of any group (esp. in the phrase **separate the sheep form the goats**). **5.** short for **scapegoat. 6. act** (or **play**) **the** (**giddy**) **goat**. to fool around. **7. get (someone's) goat**. *Slang.* to cause annoyance to (someone)

page[1] *n.* **1.** one side of one of the leaves of a book, newspaper, letter, etc. or the written or printed matter it bears. **2.** such a leaf considered as a unit **3.** an episode, phase, or period **4.** *Printing.* the type as set up for printing a page. ~*vb.* **5.** another word for **paginate. 6.** (*intr.*; foll. by *through*) to look through (a book, report, etc.); leaf through.
page[2] *n.* **1.** a boy employed to run errands, carry messages, etc., for the guests in a hotel, club, etc. **2.** a youth in attendance at official functions or ceremonies. **3.** *Medevial history.* **a.** a boy in training for knighthood in personal attendance on a knight. **b.** a youth in the personal service of a person of rank. **4.** (in the U.S.) an attendant at Congress or other legislative body. **5.** *Canadian.* a boy or girl employed in the debating chamber of the house of Commons, the Senate, or a legislative assembly to carry messages for members. ~*vb.* (*tr.*) **6.** to call out the name of (a person). **7.** to call (a person) by an electronic device, such as a bleep. **8.** to act as a page to or attend as a page.

Figure 1. Definitions of *pen, sheep, goat* and *page* in the *CED*

2.2. Connectionnist approaches

Previously suggested connectionnist models proposed by Cottrell and Small (1983) and Waltz and Pollack (1985) use a local representation of word meaning. In such representations, each node (or "neuron") represents one word or one concept. Activatory links connect words and the concepts to which they are semantically related, and lateral inhibitory links connect competing concept nodes corresponding to the various senses of a given word.[3] For example, the word *pen* is represented by a word node connected by activatory links to (at least) two concept nodes, PEN-AS-WRITING-IMPLEMENT and PEN-AS-ENCLOSURE, which are in turn connected to each other by inhibitory links. The connections in these networks and their weights are constructed manually; it is not within the scope of these studies to consider how such configurations might be learned.

These models use recurrent networks over which a relaxation process is run through a number of cycles, after the spreading activation and lateral inhibition strategy introduced by McClelland and Rumelhart (1981). Initially, the nodes corresponding to the words in the sentence to be analyzed are activated. These words activate their concept node neighbors in the next cycle, and, in turn, the neighbors activate their immediate neighbors, and so on. Concept nodes that correspond to different senses of the same word inhibit each other. These nodes can be considered to be "in competition", since the more active one sense node becomes, the more it will tend to decrease its neighbors' activation levels. The total amount of activation received by any node at any cycle is the sum of the activation and inhibition received from all of its neighbors. As the cycles progress, activation 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 network stabilizes in a configuration where the only nodes activated in the output layer are the sense nodes with the strongest relations to other nodes in the network. Because of the "winner-take-all" strategy, at most one sense node per word is ultimately activated.

Earlier experiments with neural network approaches to word sense disambiguation have demonstrated the promise of the approach. However, the models described above suffer several drawbacks. First, the networks used so far are hand-coded and thus necessarily very small (at most, a few dozen words and concepts). Due to a lack of real-size data, it is not clear that the same neural net models will scale up for realistic application. A second problem concerns the way context is handled. Early approaches rely on priming "context-setting" concept nodes to force the correct interpretation of the input words. For example, the concept node WRITING must be primed to help select the corresponding sense of *pen*. However, context-setting nodes must be artificially primed by the experimenter, and it is not clear how they could be automatically activated in a text processing system. As Waltz and Pollack (1985) point out, the word immediately corresponding to a contextually important concept (e.g., the word *writing*) may not be explicitly present in the text under analysis, although the concept can usually be inferred from other words in the text (e.g., *page, book, ink,* etc.).

To solve this problem, Waltz and Pollack (1985) and Bookman (1987) include sets of semantic "microfeatures," corresponding to fundamental semantic distinctions (animate/inanimate, edible/inedible, threatening/safe, etc.), characteristic duration of events (second, minute, hour, day, etc.), locations (city, country, continent, etc.), and other similar distinctions, in their networks. To be comprehensive, the authors suggest that these features must number in the thousands. Each concept node in the network is linked, via bidirectional activatory or inhibitory links, to a subset of the microfeatures. Collections of closely related concepts share many common microfeatures, and therefore provide a representation of context that influences the interpretation of ambiguous words in the text. In Waltz and Pollack (1985), sets of microfeatures have to be manually primed by a user, but Bookman (1987) describes a dynamic process in which the microfeatures are automatically activated by the preceding text, thus acting as a short-term context memory.

Microfeature-based schemes are problematic due to the difficulties of designing an appropriate set. Despite several efforts (see Smith and Medin, 1981), we are still very far from identifying a universally accepted list of semantic features. This becomes clear when one examines the sample microfeatures given by Waltz and Pollack: they specify microfeatures such as CASINO and CANYON, but it is obviously questionable whether such concepts constitute fundamental semantic distinctions. On a more practical level, it is simply difficult to imagine how vectors of several thousands of microfeatures for each one of the tens of thousands of words and hundreds of thousands of senses can be realistically encoded by hand.

We demonstrate that semantic microfeatures are not required in order to automatically set context from words in the preceding text. We believe that what microfeatures achieve in the models proposed by Waltz and Pollack (1985) and Bookman (1987) is a high degree of connection among concepts. We show that context can be handled automatically with a network consisting of only word and concept nodes (and no microfeatures), provided the network is large enough and very densely connected. This results in a substantially more economical model.

3. Network architecture and construction

3.1. Creating networks from dictionaries

Everyday dictionaries represent ready-made, highly connected networks of words and concepts. For example, in the *CED,* the definition of *pen* (as a writing implement) contains words such as *write, draw,* and *ink.* The definition of *page* contains *book, newspaper, letter, write, print.* The definition of *ink* contains *print, write,* and *draw*. The definition of *draw* contains *pencil* and *pen.* The definition of *book* contains *print, write, page, paper.* The definition of *paper* contains *write* and *print;* and so on. All of these connections obviously form a dense cluster of semantically related words.

Because several dictionaries are available in machine readable form, we have used them to build large networks of words and concepts. We exploit the existing structure of dictionaries, in which each word is connected to one or more senses (roughly equivalent to concepts), and each sense is in turn connected to the words in its definition. If the words *pen* and *page* are fed to such a network containing all the connections in the *CED*, we can expect that the appropriate senses of both *pen* and *page* will be triggered because of the activation they receive through their mutual, direct connections to the word *write,* as well as numerous other indirect paths (through *ink, paper,* etc.). Conversely, the sense of *pen* as enclosure, which contains words like *enclosure, domestic,* and *animal,* should receive no reinforcement and eventually die off because of the inhibition sent from the more activated sense of *pen*. The sheer density of the connections between the two related senses of *pen* and *page* should override other spurious connections (e.g., *page* → (to)*bear* → *animal* → *pen*) between other senses of the two words.

The network is constructed by a straightforward automatic procedure that does not require hand coding or sophisticated analysis of definitions. Definition texts from the *CED* are simply pre-processed to remove function words and frequent words, and all remaining words are morphologically normalized (see figures 1 and 2). A simple program then scans the pre-processed definition texts, creates the appropriate nodes and links, and assigns weights.

```
pen1.1    implement write draw ink sharp split quill metal nib
          attach holder.
pen1.2    write end implement nib.
pen1.3    style write.
pen1.4    write occupation write word.
pen1.5    long horn internal shell squid.
pen1.6    write compose.
pen2  1   enclosure domestic animal keep.
pen2.2    place confinement.
pen2.3    dock service submarine.
pen2.4    enclose keep pen.
pen3      penitentiary.
pen4      female swan.
```

Figure 2. Pre-processed definition for the word *pen*

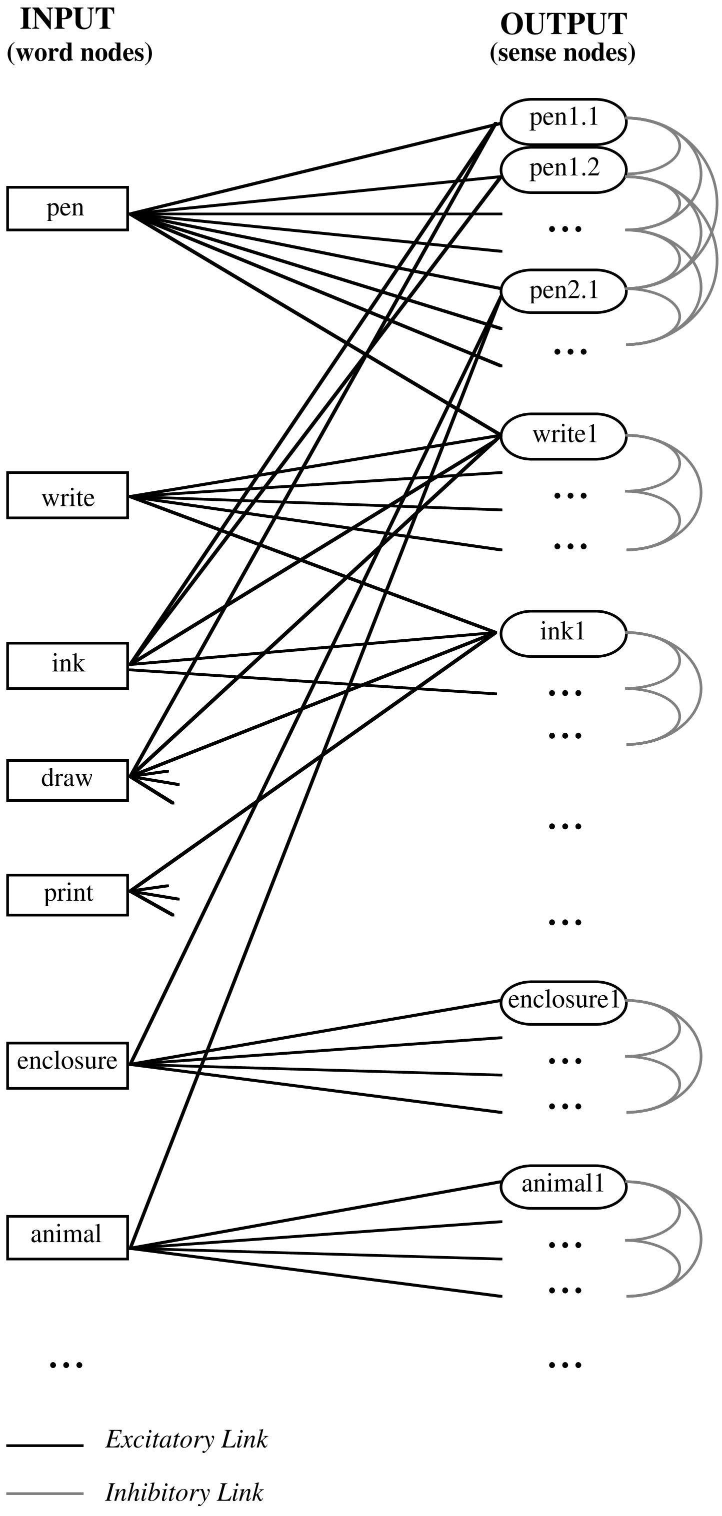

Figure 3. Topology of the network

Ideally, the network we build would include the entire dictionary. However, the resulting network for the *CED* would be enormous, and so for practical reasons we currently limit the size of the network to a few thousand nodes and 10 to 40 thousand non-null weighted links. We do this by building only the portion of the network that represents the input words, the words in their definitions, and the words in these words' definitions in turn (as well as the intervening sense nodes). Thus for each set of input words, a potentially different network is constructed.

3.2. Network architecture

Our network consists of two layers: the input layer, consisting of *word nodes*, and an output layer consisting of *sense nodes* (figure 3). Each word defined in the *CED* is represented by a word node in the input layer. For each word node a number of sense nodes in the output layer represent the different senses (definitions) of this word in the *CED.* Thus, for the lexical entry *pen*, 12 sense nodes exist, one corresponding to each of the 12 senses of *pen* given in the *CED* (figure 1). The two layers are fully interconnected: there are *feedforward* links from the input layer to the output layer, and *feedback* links from the output to the input layer (the weight on a link might be null) as well as between nodes of the output layer as lateral inhibition. This network is therefore a recurrent network (as opposed to a multilayer, feedforward network), as in McClelland and Rumelhart (1981) and Hopfield (1982, 1984). Its two-layer architecture bears some similarity to the ART network proposed by Grossberg and Carpenter (1988), in which the continually modified input vector is passed forwards and backwards (resonated) between the layers in a cyclic process.

When entries are scanned in creating the network, positive weights are assigned to links between the word node for the headword and its sense nodes, thus creating *excitatory* links. Positive weights are also assigned to links between a sense node and the nodes corresponding to the words that appear in its definition. Lateral *inhibitory* links between sense nodes of the same word node are assigned negative weights. Finally, for each connection from a node i to a node j, the same weight is assigned to the reciprocal connection from node j to node i, thus creating a symmetrically weighted network.

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

3.3. The activation strategy

We can describe the activation strategy in mathematical terms as follows. At any instant *t,* a node *i* in the network 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 towards a given node *i* from a node *j* has a weight w_{ji}, which has a positive value for excitatory links and a negative value for inhibitory links (note that for each link to node *i* from a node *j* there is a reciprocal link in the network from *i* to *j,* and $w_{ij} = w_{ji}$). At any cycle, 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" σ 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, 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 $\theta_i(a_i(t) - r_i)$:

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

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

4. Experiment

4.1. Method

Ultimately, the input to the network would be taken from a window of unanalyzed, raw text, subjected to minimal processing to remove grammatical function words and to identify root forms; or it may consist of syntactically related words (e.g., the verb and the noun-head of PP) from a partially analyzed text. However, in order to test the network, we have used a simplified case in which the input is restricted to two words only: a "context word" and a second ambiguous word. Using only two words as input enables us to better understand the behavior of the network, although with only one word of context, the disambiguation task may be more difficult. After the spreading activation algorithm is run, only one output sense node attached to each input word node is active in the network. This node should identify the *CED* sense that is intended in the given context.

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 4. A sample of the corpus

We tested this strategy with 23 clearly ambiguous words, such as *ash, bay, cape,* etc., in different contexts (figure 4). 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 *CED.* 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.

For each of the 23 words, we constructed 2 groups of 3 context words, using examples from the *Brown Corpus of American English* (Kucera and Francis, 1967) whenever possible. Each of the two groups is semantically related to one homograph. Therefore, the final experimental corpus consists of 138 word pairs.

In these experiments, we were concerned with identifying the correct homograph, rather than the correct sense within a homograph. Sense distinctions within a homograph are very often too subtle to be differentiated by the network in experiments with only one word of context, and may even be impossible for humans to distinguish. For example, the *CED* gives three senses for the second homograph of *ash:* (1) the tree, (2) the wood of that tree, and (3) any of several trees resembling the ash. In the context of the word *tree* alone, it is not clear, even for a human, which sense of *ash* is intended. Furthermore, most language processing applications would not require this level of fine semantic distinction. Because we were not interested in the exact sense selected with a homograph, we created a bias that favors the senses which appear first within each homograph (these senses are typically the most basic or common; other senses are often qualifications of the first, or metonymic uses, synecdoches, etc.--e.g., "the pen" as a style of writing). This bias is achieved with a decreasing function applied to the weights of the links between a word and its sense nodes within each homograph.

4.2. Quantitative results

The network can settle in two possible states: it can identify a disambiguated sense for the input word, or all sense nodes can "die off" after a number of cycles, thus giving no answer. We therefore need to evaluate both the *efficiency* of the method (the ratio of correct answers to the total numer of word pairs) and its *reliability* (the percentage of correct answers to the total number of actual answers). We do not expect 100% efficiency, and this method will probably have to be combined with other sources to achieve perfect results.[4] However, combining sources is possible only if each is very reliable. Therefore, it is obviously better for the network to give a "don't know" answer than to give an erroneous result.

The behavior of the network was very sensitive to parameter tuning (e.g., threshold values, decay rate, etc.). In the experiments with the best results, the correct homograph was indentified in 85% of the cases (118 word pairs), which is much better than chance (39%).[5]

In 12% of the cases (16 word pairs), the network failed to identify any homograph as the correct one, because the appropriate links between words were not found. 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 direct path between *lawn* and *house* in the *Collins English Dictionary,* although it is clear that the connection is part of human experience.

In 3% of the cases (4 word pairs) a wrong result was given because good connections exist between the wrong sense pairs. For example, the network identified the wrong sense of *ash* when given the pair *ash-tobacco,* because the connections between the "tree" sense of *ash* and *tobacco* are more direct (both are plants) than those between the "residue" sense of *ash* and *tobacco*. On logical grounds, the connection between the "tree" sense of *ash* and *tobacco* is not "wrong"; it is simply that world knowledge based on the experience of most human beings provides a stronger connection to the "residue" sense of *ash*. As in the case of *lawn-house,* this piece of knowledge is not directly represented in the dictionary.

4.3. Detailed example

Figures 5 and 6 show the state of the network after being run with *pen* and *goat,* and *pen* and *page,* respectively. The figures represent only the most activated part of each network after 100 cycles. Over the course of the run, the network reinforces only a small cluster of the most semantically relevant words and senses, and filters out the rest of the thousands of nodes. The correct sense for each word in each context (*pen* 2.1 with *goat* 1, and *pen* 1.1 with *page* 1.1) is the only one activated at the end of the run.

Sense 1.1 of *pen* would also be activated if it appeared in the context of a large number of other words--e.g., *book, ink, inkwell, pencil, paper, write, draw, sketch,* etc.--which have a similar semantic relationship to *pen*. For example, figure 7 shows the state of the network after being run with *pen* and *book*. It is apparent that the subset of nodes activated is similar to those which were activated by *page*.

4.4. Discussion

Although our model is only preliminary, the results are promising. The network's efficiency is good (85%, or 118 correct answers out of 138 cases); its reliability is also very good (97%, or 118 correct answers out of 122 actual answers).

Our results are particularly encouraging because we can see that further improvements to the network are possible without the need for substantial processing or manual encoding. We can, for instance, further analyze definition texts to identify part of speech for each word using straightforward procedures with good efficiency (see, for example, Church, 1988; DeRose, 1988). Syntactic tags are often all the information that is needed for sense discrimination, since many words which can be used in two or more syntactic categories are unambiguous within any one category. For example, the word *bear* is ambiguous only without part of speech information; distinguishing the noun *bear* and the verb *bear* eliminates the ambiguity. The identification of syntactic categories within the network would remove many spurious connections, such as the connection mentioned above between *page* and *pen* through the word *bear.*

Other kinds of information could be added to the network with low processing cost. For example, the network could be augmented to provide information about genus relations, which several studies have shown to be of special importance for language understanding and disambiguation (Fass, 1988; Wilks and Fass, 1990; Iverson and Helmreich, 1992; McRoy, 1992). In our network, links between a sense and a genus word appearing in its definition could be more heavily weighted, in order to make them bear more importance in the process of sense identification. Unlike other semantic relations, genus terms can be extracted from dictionary definitions with good accuracy using a straightforward procedure requiring no sophisticated parsing (Chodorow, Byrd, and Heidorn, 1985). Similarly, links could be created between senses of words that appear together in collocations on the basis of data extracted from corpora.

It is also possible to add weight to links between a word and its preferred sense or senses, by exploiting a source such as the *COBUILD Dictionary* (which exists in machine readable form as a part of the Vassar/CNRS database). This dictionary, constructed on the basis of word use in a 20 million word corpus of British English, lists all word senses in order of decreasing frequency of use. Domain-specific senses could also be identified using information about domain relevance such as that contained in the *Longman's Dictionary of Contemporary English* (also part of the Vassar/CNRS database). This would require the addition to the network of "domain" nodes, and/or the creation of links between word senses sharing a common domain.

Apart from adding information to the network, increasing its size can also improve its performance, since many of the failures in these experiments result from too-remote relations between the input words. As noted above, in our current implementation, the content of the network is limited to the input words, the words in their definitions, the words in the definition of these words, and the intervening sense nodes. The results may be improved with additional iterations or, ideally, with a single network covering the entire dictionary. For example, in the case of *lawn-house, garden* is not included in the network. However, since the definition of *garden* contains *house,* and at the same time shares the words *grass* and *area* with the definition of *lawn,* there would be a straightforward path between *lawn* and *house* if the entire dictionary were used. We are currently working on a large-scale implementation which would enable us to experiment with the use of a network built from an entire dictionary.

In general, our results show that a connectionist model embodying information about word associations can contribute significantly to the process of automatic word sense disambiguation. Our model and method seem to at least partially overcome many of the shortcomings of the word association/semantic context approach, often criticized for revealing too many false paths between words or concepts, because of the "winner-take-all" strategy and the ability to adjust link weights and network parameters. Nonetheless, we realize that the approach has its limitations, and with McRoy, we believe that the best strategy will ultimately prove to to be one that utilizes multiple knowledge sources. However, we are more optimistic than McRoy about the possibilities for automatic word sense disambiguation without the cost of creating massive semantic knowledge bases. McRoy herself states that their studies show the most important sources of information for word sense disambiguation are syntactic tags, morphology, collocations, and word associations. Our network currently includes two of these four sources of information (morphology and word associations), and could be augmented to include syntactic tags with little processing cost. We expect that the network with some augmentation of this kind could perform significantly better.

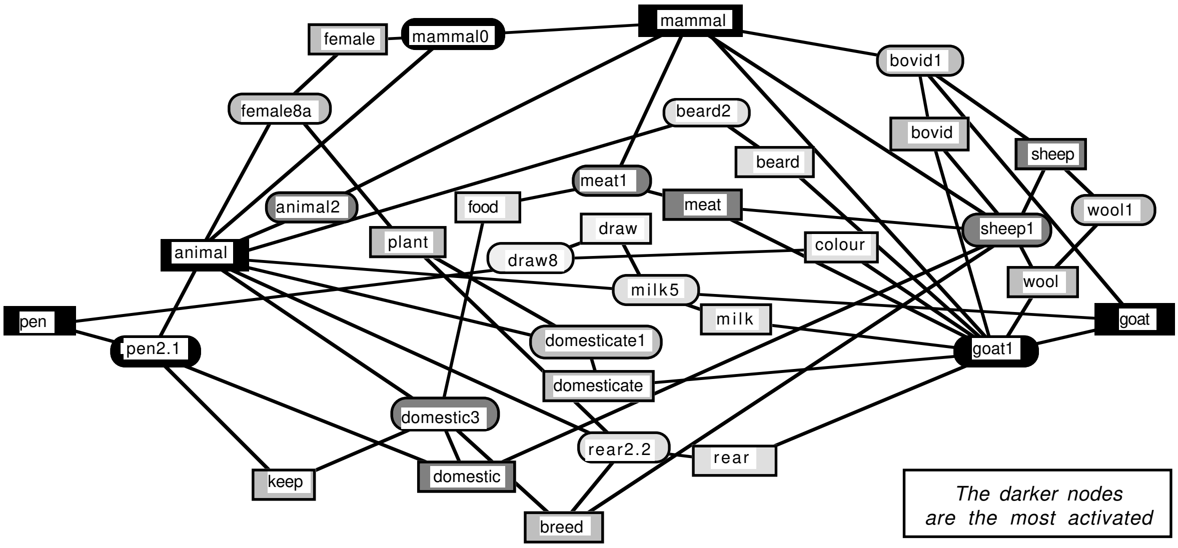

Figure 5. State of the network after being run with *pen* and *goat*

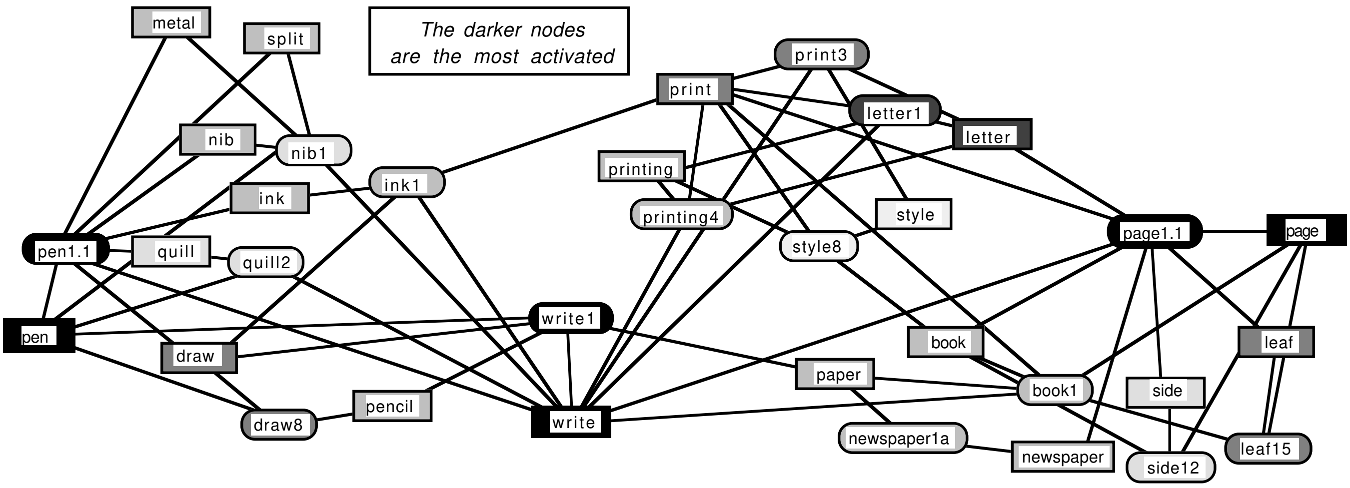

Figure 6. State of the network after being run with *pen* and *page*

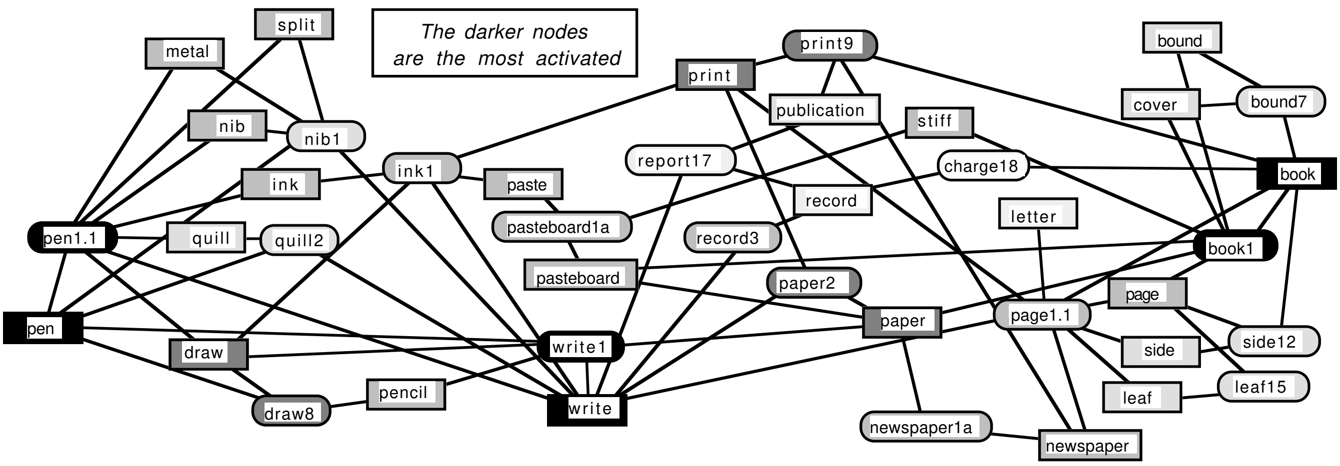

Figure 7. State of the network after being run with *pen* and *book*

5. Conclusion

We construct networks of words and concepts from machine readable dictionaries in order to achieve word sense disambiguation. This enables real-size experiments with models proposed previously (Cottrell and Small, 1983; Waltz and Pollack, 1985; Bookman, 1987). The good success rate in our results demonstrates the validity of the approach, and suggests that used in conjunction with other knowledge sources, it can be used effectively to disambiguate words in unrestricted text. We also show that there is no need for an additional layer of microfeatures to provide contextual information, as proposed in Waltz and Pollack (1985) and Bookman (1987). The high density of connections in a network build from dictionaries can provide adequate contextual information, especially if the entire dictionary were to be represented in the network.

Notes

[1] For example, preliminary evidence shows that sense tagging of words in the input text significantly improves the performance of information retrieval systems (Krovetz, 1989).

[2] This partially explains why a recent survey of existing language processing systems (excluding large machine translation systems) revealed that the average size of their lexicons is only 30 or 40 words (Boguraev and Briscoe, 1989).

[3] These models integrate syntactic and semantic information in the network. Here, we are concerned only with the semantic component.

[4] The use of multiple knowledge sources for word sense disambiguation is increasingly recognized as a promising strategy (McRoy, 1992).

[5] Note that these results are better than those given in Ide and Véronis (1990), due to the discovery of better parameter values in later experiments.

Acknowledgements

The present research has been partially funded by the GRECO-PRC Communication Homme-Machine of the French Ministery of Research and Technology, U.S.-French NSF/CNRS grant INT-9016554 for collaborative research, and U.S. NSF RUI grant IRI-9108363. The authors would like to thank Collins Publishers for making their data available for research within the project. The authors would also like to acknowledge the contribution of Stéphane Harié for his pre-processing of the *CED*.

References

AMSLER, R. A. (1980). *The structure of the Merriam-Webster Pocket Dictionary.* Ph. D. Dissertation, University of Texas at Austin.

BOGURAEV, B., BRISCOE, T. (1989). *Computational Lexicography for Natural Language Processing*. London and New York: Longman.

BOOKMAN, L.A. (1987). A Microfeature Based Scheme for Modelling Semantics. *Proceedings of the 10th International Joint Conference on Artificial Intelligence, IJCAI'87,* Milan, Italy, 611-614.

BYRD, R. J., CALZOLARI, N., CHODOROV, M. S., KLAVANS, J. L., NEFF, M. S., RIZK, O. (1987) Tools and methods for computational linguistics. *Computational Linguistics,* 13, 3/4, 219-240.

CALZOLARI, N.(1984). Detecting patterns in a lexical data base. *COLING'84,* 170-173.

CHARNIAK, E. (1983). Passing markers: A theory of contextual influence in language comprehension. *Cognitive Science,* 7, 3, 171-90.

CHODOROW, M. S., BYRD. R. J., HEIDORN, G. E. (1985). Extracting semantic hierarchies from a large on-line dictionary. *ACL Conference Proceedings,* 299-304.

CHURCH, K. W. (1988). A stochastic parts program and noun phrase parser for unrestricted texts. In *Proceedings of the Second Conference on Applied Natural Language Processing.* Austin, Texas, 136-143.

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

DEROSE, S. J. (1988) Grammatical category disambiguation by statistical optimization. *Computational Linguistics,* 14, 1, 31-39.

FASS, D. (1988). Collative semantics: A semantics for natural language processing. MCCS-88-118, Computing Research Laboratory, New Mexico State University, Las Cruces, NM.

FILLMORE, C. (1968). The case for case. In Bach and Harms (Eds.), *Universals in Linguistic Theory,* Holt, Chicago, IL.

GALE, W., CHURCH, K., YAROWSKY, D. (forthcoming, 1993). A method for disambiguating word senses in a large corpus. *Computers and the Humanities,* Special issue on Common Methodologies in Computational Linguistics and Humanities Computing, N. Ide and D. Walker, eds.

GROSSBERG AND CARPENTER (1988). The ART of Adaptive Pattern Recognition. *IEEE Computer,* 21, 3.

HALLIDAY, M., HASAN, R. (1976). *Cohesion in English.* Longman, London.

HIRST, G. (1987). *Semantic Interpretation and the Resolution of Ambiguity.* Cambridge University Press, Cambridge.

HOPFIELD, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences, U.S.A.,* 79, 2554-2558.

HOPFIELD, J. J. (1984). Neurons with graded response have collective computational properties like those of two-state neurons. *Proceedings of the National Academy of Sciences, U.S.A.,* 81, 3088-3092.

IDE, N.M., VÉRONIS, J. (1990). Very large neural networks for word sense disambiguation. *Proceedings of the 9th European Conference on Artificial Intelligence, ECAI'90,* Stockholm , 366-368.

IVERSON, E., HELMREICH, S. (1992). Metallel: An integrated approach to non-literal phrase interpretation. *Computational Intelligence,* 8, 1, 477-93.

KLAVANS, J., CHODOROW, M., WACHOLDER, N (1990). From dictionary to knowledge base via taxonomy. *Proceedings of the 6th Conference of the UW Centre for the New OED,* Waterloo, 110-132.

KROVETZ, R. (1989). Lexical acquisition and information retrieval. *First International Lexical Acquisition Workshop,* Uri Zernick, ed.

KUCERA, H., NELSON, F. W. (1967). Computational Analysis of Present-Day American English. Brown University Press, Providence, RI.

LESK, M. (1986). Automated Sense Disambiguation Using Machine-readable Dictionaries: How to Tell a Pine Cone from an Ice Cream Cone. *Proceedings of the 1986 SIGDOC Conference.*

MARKOWITZ, J., AHLSWEDE, T., EVENS, M. (1986). Semantically significant patterns in dictionary definitions. *ACL Conference Proceedings,* 112-119.

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.

MCROY, S. W. (1992). Using Multiple knowledge Sources for Word Sense Discrimination. *Computational Linguistics,* 18, 1, 1-30.

MORRIS, J., HIRST, G. (1991). Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational Linguistics,* 17, 1, 21-48.

NAKAMURA, J., NAGAO, M. (1988). Extraction of semantic information from an ordinary English dictionary and its evaluation. *COLING'88,* 459-464.

NORVIG, P. (1989). Marker passing as a weak method for text inferencing. *Cognitive Science,* 13, 4, 569-620.

SCHANK, R. C. (1975). *Conceptual Information Processing.* North Holland, Amsterdam.

SMITH, E. E., MEDIN, D. L. (1981). *Categories and concepts.* Harvard University Press, Cambridge, MA.

VÉRONIS, J., IDE, N.M. (1991). An assessment of information automatically extracted from machine readable dictionaries, *Proceedings of the Fifth Conference of the European Chapter of the Association for Computational Linguistics,* Berlin, Germany, 227-232.

WALKER, D.E., AMSLER, R.A. (1986). The use of machine-readable dictionaries in sublanguage analysis. In R. GRISHMAN and R. KITTEDGE (Eds.). *Analysing Language in restricted domains,* Lawrence Erlbaum, Hillsdale, NJ.

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

WILKS, Y., D. FASS (1990). Preference semantics: A family history. MCCS-90-194, Computing Research Laboratory, New Mexico State University, Las Cruces, NM.

WILKS, Y., D. FASS, C. GUO, J. MACDONALD, T. PLATE, B. SLATOR (1990). Providing Machine Tractable Dictionary Tools. In J. PUSTEOVSKY (Ed.), *Semantics and the Lexicon.* MIT Press, Cambridge, MA.