Word Sense Disambiguation with Very Large Neural Networks Extracted from Machine Readable Dictionaries

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

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

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

Abstract

In this paper, we describe a means for automatically building very large neural networks (VLNNs) from definition texts in machine-readable dictionaries, and demonstrate the use of these networks for word sense disambiguation. Our method brings together two earlier, independent approaches to word sense disambiguation: the use of machine-readable dictionaries and spreading and activation models. The automatic construction of VLNNs enables real-size experiments with neural networks for natural language processing, which in turn provides insight into their behavior and design and can lead to possible improvements.

1. Introduction

Automated language understanding requires the determination of the concept which a given use of a word represents, a process referred to as *word sense disambiguation* (WSD). WSD is typically effected in natural language processing systems by utilizing semantic feature lists for each word in the system's lexicon, together with restriction mechanisms such as case role selection. However, it is often impractical to manually encode such information, especially for generalized text where the variety and meaning of words is potentially unrestricted. Furthermore, restriction mechanisms usually operate within a single sentence, and thus the broader context cannot assist in the disambiguation process.

In this paper, we describe a means for automatically building Very Large Neural Networks (VLNNs) from definition texts in machine-readable dictionaries, and demonstrate the use of these networks for WSD. Our method brings together two earlier, independent approaches to WSD: the use of machine-readable dictionaries and spreading and activation models. The automatic construction of VLNNs enables real-size experiments with neural networks, which in turn provides insight into their behavior and design and can lead to possible improvements.

2. Previous work

2.1. Machine-readable dictionaries for WSD

There have been several attempts to exploit the information in machine-readable versions of everyday dictionaries (see, for instance, Amsler, 1980; Calzolari, 1984; Chodorow, Byrd and Heidorn, 1985; Markowitz, Ahlswede and Evens, 1986; Byrd *et al.*, 1987; Véronis, Ide and Wurbel, 1989), in which an enormous amount of lexical and semantic knowledge is already "encoded". 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 even in its base form, information from machine-readable dictionaries can be used, for example, to assist in the disambiguation of prepositional phrase attachment (Jensen and Binot, 1987), or to find subject domains in texts (Walker and Amsler, 1986).

The most general and well-known attempt to utilize information in machine-readable dictionaries for WSD 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

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

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,* the dictionary used in our experiments, in figure 1.

Figure 1: Definitions of PEN, SHEEP, GOAT and PAGE in the Collins English Dictionary

pen^1 **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. **6.** to write or compose.
pen^2 **1.** an enclosure in which domestic animals are kept. **2.**any place of confinement. **3.** a dock for servicing submarines. **4.** to enclose or keep in a pen.
pen^3 short for **penitentiary**.
pen^4 *a female swan.*

sheep 1. any of various bovid mammals of the genus *Ovis* and related genera 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. 3.** a meek or timid person. **4. separate the sheep from the goats.** to pick out the members of any group who are *superior in some respects.*

goat 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. **3.** a lecherous man. **4.** a bad or inferior member of any group **6. act** (or play) **the** (giddy) **goat.** to fool around. **7. get** (someone's) **goat.** to cause annoyance to (someone)

page^1 **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. **6.** to look through (a book, report, etc.); leaf through.
page^2 **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. 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.** an attendant at Congress or other legislative body. **5.** 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. **6.** to call out the name of (a person). **7.** to call (a person) by an electronic device, such as bleep. **8.** to act as a page to or attend as a page.

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.

2.2. Neural networks for WSD

Neural network approaches to WSD have been suggested (Cottrell and Small, 1983; Waltz and Pollack, 1985). These models consist of networks in which the nodes ("neurons") represent words or concepts, connected by "activatory" links: the words activate the concepts to which they are semantically related, and vice versa. In addition, "lateral" inhibitory links usually interconnect competing senses of a given word. Initially, the nodes corresponding to the words in the sentence to be analyzed are activated. These words activate their neighbors in the next cycle in turn, these neighbors activate their immediate neighbors, and so on. After a number of cycles, the network stabilizes in a state in which one sense for each input word is more activated than the others, using a parallel, analog, relaxation process.

Neural network approaches to WSD seem able to capture most of what cannot be handled by overlap strategies such as Lesk's. However, the networks used in experiments 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. Further, some approaches rely on "context-setting" nodes to prime particular word senses in order to force the correct interpretation. But as Waltz and Pollack point out, it is possible that such words (e.g., *writing* in the context of *pen*) are not explicitly present in the text under analysis, but may be inferred by the reader from the presence of other, related words (e.g., *page, book, inkwell,* etc.). To solve this problem, words in such networks have been represented by sets of semantic "microfeatures" (Waltz and Pollack, 1985; Bookman, 1987) which correspond 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 that humans typically make about situations in the world. To be comprehensive, the authors suggest that these features must number in the thousands. Each concept in the network is linked, via bidirectional activatory or inhibitory links, to only a subset of the complete microfeature set. A given concept theoretically shares several microfeatures with concepts to which it is closely related, and will therefore activate the nodes corresponding to closely related concepts when it is activated itself.

However, such schemes are problematic due to the difficulties of designing an appropriate set of microfeatures, which in essence consists of designing semantic primitives. 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. More practically, 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.

3. Word sense disambiguation with VLNNs

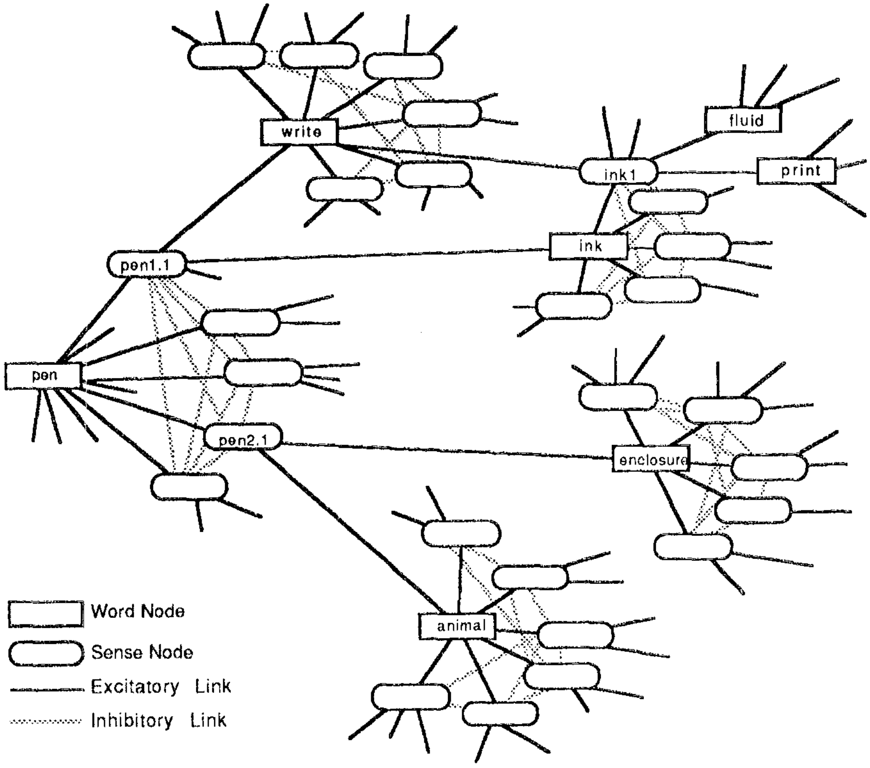
Our approach to WSD takes advantage of both strategies outlined above, but enables us to address solutions to their shortcomings. This work has been carried out in the context of a joint project of Vassar College and the Groupe Représentation et Traitement des Connaissances of the Centre National de la Recherche Scientifique (CNRS), which is concerned with the construction and exploitation of a large lexical data base of English and French. At present, the Vassar/CNRS data base includes, through the courtesy of several editors and research institutions, several English and French dictionaries (the *Collins English Dictionary*, the *Oxford Advanced Learner's Dictionary*, the *COBUILD Dictionary*, the *Longman's Dictionary of Contemporary English*, the*Webster's 9th Dictionary*, and the *ZYZOMYS* CD-ROM dictionary from Hachette Publishers) as well as several other lexical and textual materials (the *Brown Corpus of American English*, the CNRS *BDLex* data base, the *MRC Psycholinguistic Data Base*, etc.).

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

3.1. Topology of the network

In our model, words are complex units. Each word in the input is represented by a *word node* connected by excitatory links to *sense nodes* (figure 2) representing the different possible senses for that word in the *Collins English Dictionary*. Each sense node is in turn connected by excitatory links to word nodes representing the words in the definition of that sense. This process is repeated a number of times, creating an increasingly complex and interconnected network. Ideally, the network would include the entire dictionary, but for practical reasons we limit the number of repetitions 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. The different sense nodes for a given word are interconnected by lateral inhibitory links.

Figure 2. Topology 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 receive *feedback* from connected nodes. Competing sense nodes send inhibition to 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 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.

Our model does not use microfeatures, because, as we will show below, the context is taken into account by the number of nodes in the network and the extent to which they are heavily interconnected. So far, we do not consider the syntax of the input sentence, in order to focus on the semantic properties of the model. However, it is clear that syntactic information can assist in the disambiguation process in certain cases, and a network including a syntactic layer, such as that proposed by Waltz and Pollack, would undoubtedly enhance the model's behavior.

3.2. *Results*

The network finds the correct sense in cases where Lesk's strategy succeeds. For example, if the input consists of *pen* and *sheep*, *pen* 2.1 and *sheep* 1 are correctly activated. More interestingly, the network selects the appropriate senses in cases where Lesk's strategy fails. Figures 3 and 4 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.

This model solves the context-setting problem mentioned above without any use of microfeatures. 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 5 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*.

Figure 3. State of the network after being run with "pen" and "goat"

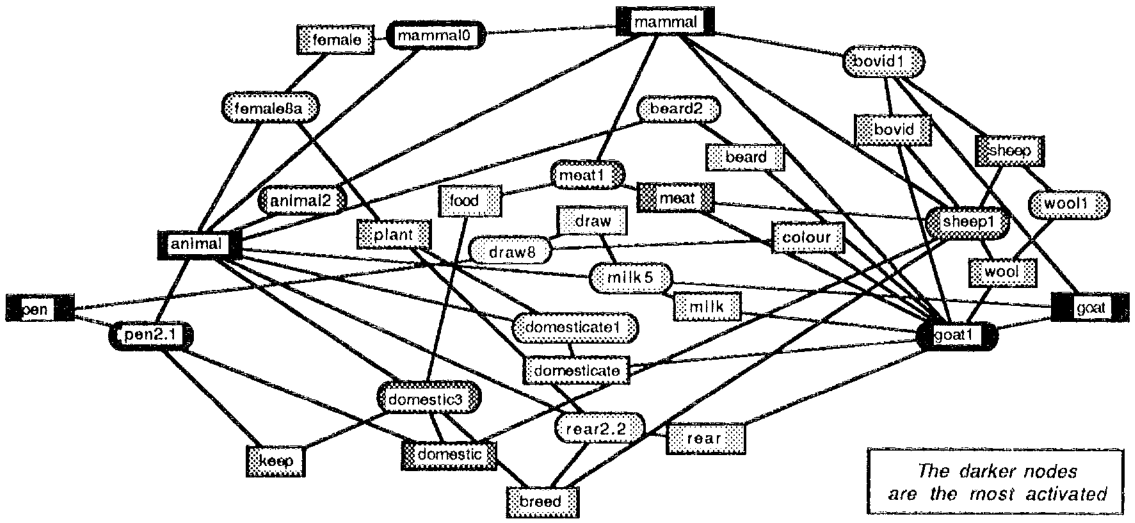


Figure 4. State of the network after being run with "pen" and "page"

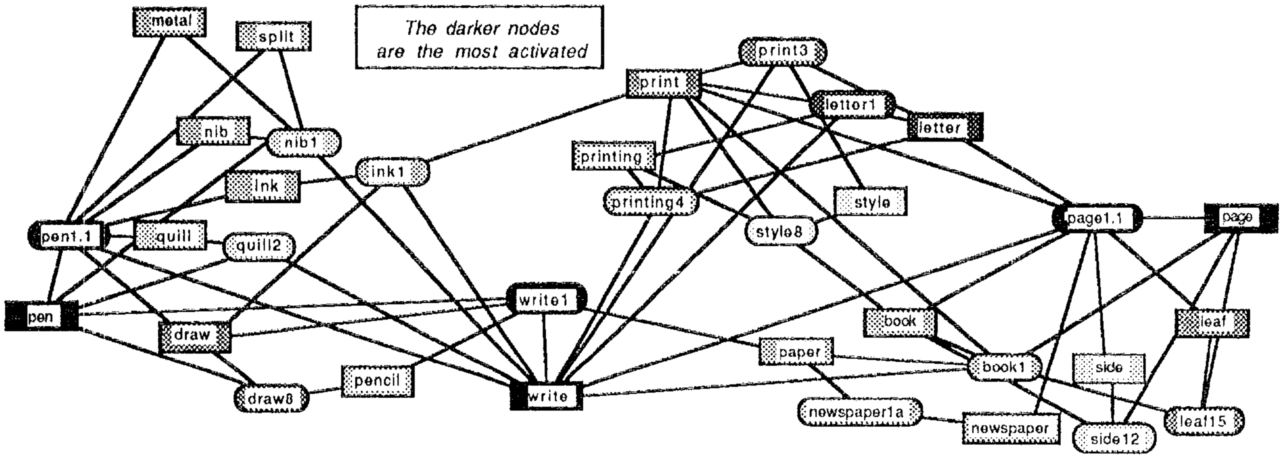
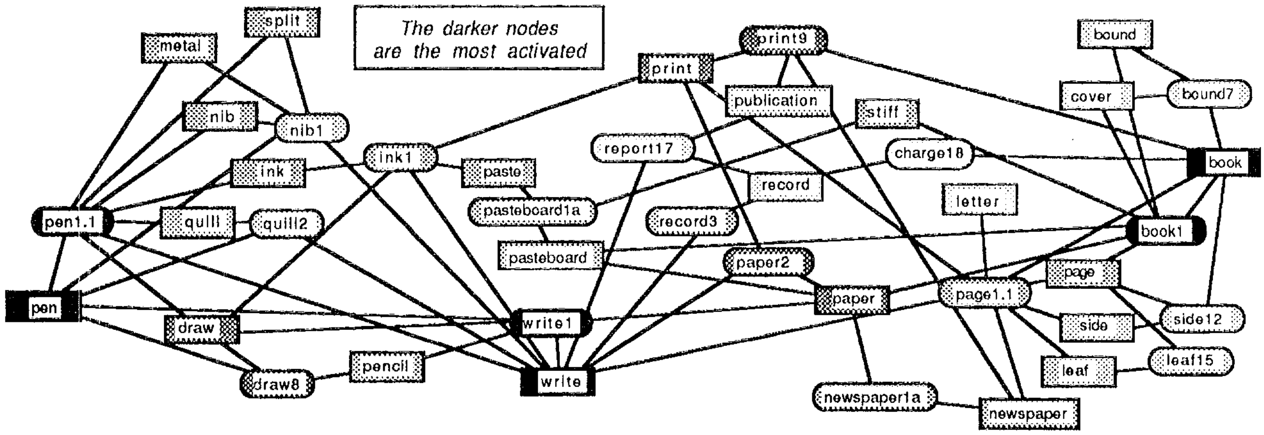


Figure 5. State of the network after being run with "pen" and "book"



The examples given here utilize only two words as input, in order to show clearly the behavior of the network. In fact, the performance of the network improves with additional input, since additional context can only contribute more to the disambiguation process. For example, given the sentence *The young page put the sheep in the pen*, the network correctly chooses the correct senses of *page* (2.3: "a youth in personal service"), *sheep* (1), and *pen* (2.1). This example is particularly difficult, because *page* and *sheep* compete against each other to activate different senses of *pen*, as demonstrated in the examples above. However, the word *young* reinforces sense 2.3 of *page*, which enables *sheep* to win the struggle. Inter-sentential context could be used as well, by retaining the most activated nodes within the network during subsequent runs.

By running various experiments on VLNNs, we have discovered that when the simple models proposed so far are scaled up, several improvements are necessary. We have, for instance, 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 addition, we have found that is is necessary to treat a word node and its sense nodes as a complex, ecological unit rather than as separate entities. In our model, word nodes control the behavior of sense nodes by means of a differential neuron that prevents, for example, a sense node from becoming more activated than its master word node. Our experimentation with VLNNs has also shed light on the role of and need for various other parameters, such as thresholds, decay, etc.

4. Conclusion

The use of word relations implicitly encoded in machine-readable dictionaries, coupled with the neural network strategy, seems to offer a promising approach to WSD. This approach succeeds where the Lesk strategy fails, and it does not require determining and encoding microfeatures or other semantic information. The model is also more robust than the Lesk strategy, since it does not rely on the presence or absence of a particular word or words and can filter out some degree of "noise" (such as inclusion of some wrong lemmas due to lack of information about part-of-speech or occasional activation of misleading homographs). However, there are clearly several improvements which can be made: for instance, the part-of-speech for input words and words in definitions can be used to extract only the correct lemmas from the dictionary, the frequency of use for particular senses of each word can be used to help choose among competing senses, and additional knowledge can be extracted from other dictionaries and thesauri. It is also conceivable that the network could "learn" by giving more weight to links which have been heavily activated over numerous runs on large samples of text. The model we describe here is only a first step toward a fuller understanding and refinement of the use of VLNNs for language processing, and it opens several interesting avenues for further application and research.

References

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

BOOKMAN, L.A. (1987). A Microfeature Based Scheme for Modelling Semantics. *Proc. IJCAI'87*, Milan, Italy, 611-14.

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.

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

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

JENSEN, K., BINOT, J.-L. (1987). Disambiguating prepositional phrases by using on-line dictionary definitions. *Computational Linguistics*, **13**, 3/4, 251-260.

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

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

VÉRONIS, J., IDE, N.M., WURBEL, N. (1989). Extraction d'informations sémantiques dans les dictionnaires courants, *7ème Congrès Reconnaissance des Formes et Intelligence Artificielle*, AFCET, Paris, 1381-1395.

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, C. GUO, J. MACDONALD, T. PLATE, B. SLATOR (forthcoming). Providing Machine Tractable Dictionary Tools. In J. PUSTEOVSKY (Ed.), *Theoretical and Computational Issues in Lexical Semantics*.