If You Have It, Flaunt It: Using Full Ontological Knowledge for Word Sense Disambiguation[1]

Kavi Mahesh, Sergei Nirenburg and Stephen Beale
Computing Research Laboratory (3CRL)
New Mexico State University
Las Cruces, NM 88003, USA
{mahesh,sergei,sb}@crl.nmsu.edu
http://crl.nmsu.edu/Research/Projects/mikro/

Abstract. Word sense disambiguation continues to be a difficult problem in natural language processing. Current methods, such as marker passing and spreading activation, for applying world knowledge in the form of selectional preferences to solve this problem do not make effective use of available knowledge. Moreover, their effectiveness decreases as the knowledge is made richer by acquiring more and more conceptual relationships. Effective resolution of word sense ambiguities requires inferring the dynamic context in processing a sentence in order to find the right selectional preferences to be applied. In this article, we propose such an inference operator and show how it finds the most specific context to resolve word sense ambiguities in the Mikrokosmos semantic analyzer. Our method retains its effectiveness even in a rich, large-scale knowledge base with a high degree of connectivity among its concepts.

1. Disambiguation in Context

Word sense disambiguation continues to be a difficult problem for programs that process natural language. The goals of word sense resolution methods are: (a) to select as small a subset of possible senses of a word as possible, ideally just one sense, and (b) to select the best sense(s) given all the knowledge available to the system, including the dynamic context in processing the text. The most common methods for resolving word sense ambiguities are based on statistical collocations or selectional preferences (for a recent survey, see Guthrie et al, 1996) between pairs of word senses. Often, individual selectional preferences applicable to a word are not strong enough to exclude all but one sense of the word. The real power of word sense selection seems to lie in the ability to constrain the possible senses of a word based on selections made for other words in the dynamic context.

Although it is a truism that context plays a significant role in sense disambiguation, computational models have not demonstrated the effectiveness of modeling context for resolving word senses in a large-scale NLP system.

This work presupposes a semantic analysis environment, such as the Mikrokosmos system (Beale et al, 1995; Mahesh et al, submitted; Onyshkevych and Nirenburg, 1995), in which the results are expressions in a text meaning representation language whose syntax is based on propositions and their arguments, as well

1. Research reported in this paper was supported in part by Contract MDA904-92-C-5189 from the U.S. Department of Defense.

as information about speaker attitudes and the speech situation. The vocabulary of this language is defined by a large ontological model of the world, and the lexicon formulates selectional restrictions and other semantic constraints in terms of ontological concepts. The focus of this paper is on control issues in applying selectional preferences to achieve effective resolution of sense ambiguities. Word sense disambiguation in a basic semantic analysis system is addressed in a parallel submission (Mahesh et al, submitted). In this paper, we argue that:

- individual constraints between the head of a proposition and each of its arguments typically available in static knowledge sources (lexicons) are often not strong enough for effective selection of word senses; knowledge of constraints and conceptual relationships among the arguments of a proposition is often critical;
- effective sense disambiguation requires rich knowledge with a high degree of cross-dependence among knowledge elements;
- it is often not possible to determine a diagnostic context statically (i.e., before any decisions are made for the current sentence);
- while representations such as semantic networks (including both simple labeled hierarchies (e.g., SENSUS (Knight and Luk, 1994) and ontological concept networks (e.g., the Mikrokosmos ontology (Mahesh, 1996; Mahesh and Nirenburg, 1995)) can capture such constraints and relationships, processing methods currently applied to semantic networks such as marker passing (Charniak, 1983; 1986; Eiselt, 1989) and spreading activation (Waltz and Pollack, 1985) do not facilitate selection of word senses based on the dynamic context;
- marker passing and spreading activation are effective on well-designed and sparse networks but become less and less effective as the degree of connectivity increases.

2. Knowledge Requirements for Word Sense Disambiguation

If one chooses a semantic network-like formalism to encode large amounts of highly interdependent knowledge elements required for a processing task, the network will exhibit a high degree of connectivity. In NLP, more connections means more knowledge that is potentially useful in finding semantic connections between the different parts of a sentence or text. We will describe the processing in our approach using a simple (but un-contrived) example.

Consider a sentence such as “John prepared a cake with the range.” Leaving aside, for the sake of simplicity, the PP-attachment ambiguity, let us concentrate on lexical disambiguation. In this sentence, several words are ambiguous, relative to the static knowledge sources we used in our experiment: “Prepared” can mean *prepare-food*[1] or *prepare-document,* “cake” is A-KIND-OF *baked-food* and also A-KIND-OF *dessert,* and “range” can mean *oven* or *stove,* in addition to the mathematical, military and agricultural senses.

Suppose that the analysis process has already ruled out, based on static context, the non-kitchen senses of “range” and determined the correct sense of “prepared.” The ontological concept *prepare-food* has *prepared-food* as its THEME; *human* as its AGENT; and *cooking-equipment* as its INSTRUMENT. It also has *bake* as one of its descendants. *Bake* is constrained as follows: its INSTRUMENT is *oven;* its AGENT is *baker* (this constraint is made RELAXABLE-TO *human,* as a preparation for processing non-literal language — see Ony-

1. Words in double quotes “” denote lexemes; words in *italics* denote ontological concepts, words in SMALL CAPS denote properties (and properties of properties) of ontological concepts.

shkevych and Nirenburg (1995) for details); its THEME is *baked-food* (of which, recall, *cake* is A-KIND-OF).

One cannot realistically expect an English lexicon to contain a selectional constraint associated with the INSTRUMENT role of *prepare-food* that enables the system to distinguish between *oven* and *stove,* both direct ontological descendants of *cooking-equipment,* because any kind of cooking-equipment can be the instrument of preparing food. However, as soon as it is determined that the food in question is a cake, which is a kind of baked food, the kinds of cooking equipment that can used will become further constrained. In particular, *stove* will be ruled out, leaving *oven* as the only remaining sense of "range."

How can this dynamic constraint on *cooking-equipment* be introduced? In the ideal situation, there will be a direct connection in the network between *oven* and *baked-food* but none between *stove* and *baked-food* and the system selects oven based on this information. Such inter-argument, "lateral" selectional restrictions (i.e., those not anchored at the head of a proposition, such as *bake,* in this case) seem to be invaluable for word sense disambiguation.

In reality, however, things are more complicated and the availability of such links cannot be guaranteed. As a result, NLP systems that depend on always having such information have not been highly successful in domain-independent word sense disambiguation. It would be safe to assume that knowledge sources for NLP are always incomplete and inaccurate, due to limitations of manual acquisition. The necessary links (such as the *oven* to *baked-food* link) may not have been acquired. On the other hand, some other links may be present between *stove* and *baked-food,* which might complicate the decision if the measure of semantic distance used in disambiguation is based only on connectivity (see Mahesh et al (submitted) for a discussion of the ontological search algorithm for determining semantic distance between any two ontological concepts).

We show how the dynamic context helps resolve the ambiguity even in the absence of complete knowledge (such as a direct connection between *oven* and *baked-food.*). Then we show why spreading activation methods fail to apply available knowledge effectively and how the method we propose determines the dynamic context correctly and applies the same knowledge effectively to select the oven sense of "range," even in the absence of a direct (or "short," "low-cost") link between *oven* and *baked-food.* Then we show how the proposed method retains its effectiveness even at higher degrees of connectivity while marker passing and spreading activation methods lose their disambiguation power as more knowledge is introduced.

3. Dynamic Context Helps Select the Right Sense

An important point to note from the above example is that *bake* was not explicitly referred to in the sentence. Nevertheless, once "cake" is determined to be *baked-food,* the processor should be able to infer that the meaning of "prepared" must have, in fact, been *bake* since that is the only subclass of *prepare-food* that takes *baked-food* as THEME. (In other words, the system, in fact, "corrects" the author of the input text, suggesting that in English "bake" should have been used in this sentence.) This information is included in the dynamic context only after it is determined that "cake" refers to baked-food. Once this dynamic context is inferred, constraints are modified; in this case, tightened. As shown in Figure 1, we know that the INSTRUMENT of *bake* is constrained to be an instance of *oven,* and hence that is the appropriate sense of "range," while the instrument of *prepare-food* could be any *cooking-equipment.* By tracking the dynamic context, we can exploit such constraints between two (or more) arguments (the THEME *baked-food* and the INSTRU-

MENT *oven*) of a head *(prepare-food)* even when there is no direct link between the arguments.

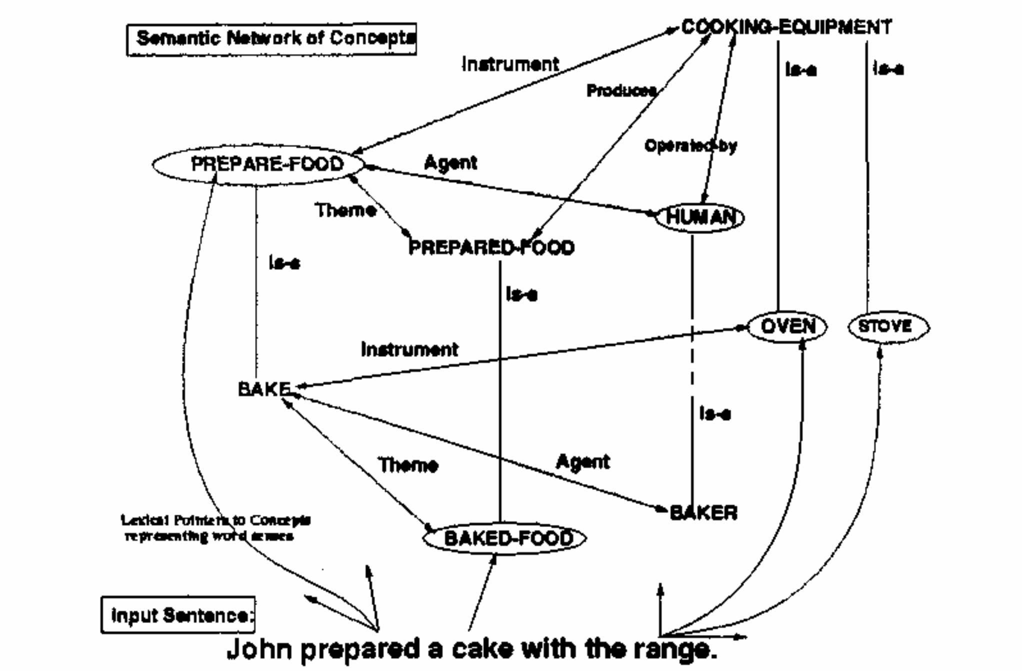

Figure 1. A Semantic Network.

Current methods based on constraint satisfaction techniques do not make this inference and hence fail to apply the stronger constraints available in the system's knowledge bases. They do not necessarily apply constraints attached to intermediate nodes, that is, those nodes in the conceptual network that are along the paths between origin nodes pointed at by the lexical entries for the words in the sentence. Such constraints seem to play a critical role in word sense disambiguation. For example, in the sentence above, it is not possible to discriminate between the senses of “range” without looking at the constraints attached to the intermediate node *bake.*

Note that the problem is one of control of finding the right constraints to apply rather than the correctness of propagating directly available constraints. Can marker passing or spreading activation accomplish this? Yes, but only by guessing the dynamic context with the help of heuristics based on the topology of the network, not the content of the knowledge in the network. The methods are too weak to guarantee that the guessed context is the right one given all available knowledge. This is because the methods are unduly influenced by knowledge in the network that is not relevant to the current context. The following section illustrates how well-intentioned links among other related concepts can unduly influence the methods so as not to include *bake* in the current context.

4. Spreading Markers or Activation: A Game of Luck

In the case of marker passing, there may be many other paths of the same or shorter length connecting pairs of concepts that do not go through nodes in the current context such as *bake.* In Figure 1, for example, there is an alternative path from *baked-food* to *prepare-food* via *prepared-food.* This path consists of a

THEME segment and an IS-A segment just as the one going through *bake.* Thus, any choice in a marker passing algorithm will be hampered.

Let us assume that "John" is an instance of *human.* The following nodes become the origins for marker passing: *human, prepare-food,* other senses of "prepare", *baked-food, oven, stove* and other senses of "range." The goal of marker passing is to find a shortest path between each pair of origins. In pure marker passing, there are no weights on links; they carry a unit cost. Some candidates for shortest paths are shown in Figure 2.

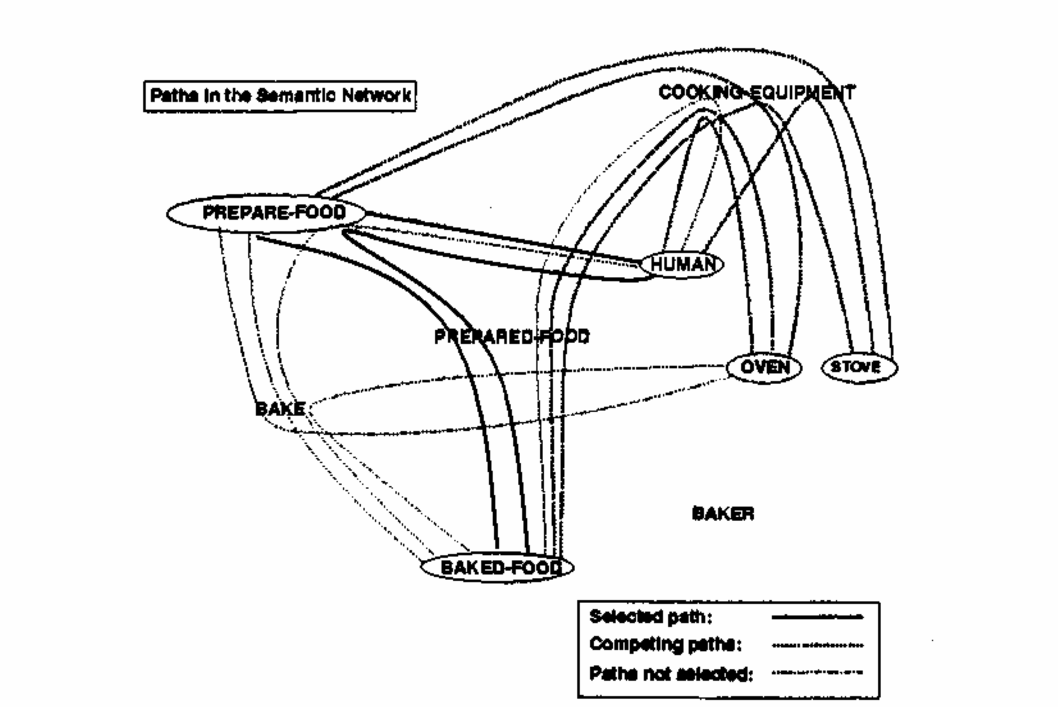

Figure 2: Paths in the Semantic Network.

Paths are selected from the set in Figure 2 by applying several heuristics such as finding a minimal spanning tree for the origin set or maximizing the set of shared nodes in the paths (Eiselt, 1989). It is clear from the figure that *cooking-equipment* and *prepared-food* are strong intermediate nodes. *Bake* might lose against these two and if so, the path from *oven* to *baked-food* via *bake* may be rejected and the competing path via *prepared-food* selected in order to maximize measures such as the total number of shared nodes among the selected paths.

As a result, *oven* and *stove* turn out to be equally likely! Although *bake* had created a shorter path between *oven* and *baked-food* than between *stove* and *baked-food,* other parts of the network had an undue advantage over *bake* as a result of well-intentioned heuristics such as the above. In this situation, it is only by luck that *oven* might get selected, or that the heuristics discriminate between competing word senses sufficiently for any selection to take place at all.

Knowledge of *bake* which was clearly present and accessed during processing was not applied effectively by the marker passing method to make the right selection for the sense of "range." Similarly, in spreading activation, *prepared-food* and *cooking-equipment* receive a high amount of activation, once again jeopardizing the role of *bake* in selecting oven rather than stove. There is no guarantee that the configuration in the network will result in a higher activation of *oven* than other senses because of the activation flowing through *bake.*

Figure 1 shows a small fragment of a conceptual network, with only a few types of available links listed. Any realistic model will have a much larger network with many other types of links between concepts, further decreasing the chances that the desired path through *bake* will be the least-cost path in the context of a sentence such as the one above. Moreover, these networks are almost always hand-coded and may include spurious links that eventually bypass certain desired paths. Processing mechanisms such as marker passing and spreading activation are simple and have a cognitive appeal, but their lack of reference to the content of the nodes makes them too weak for making the kinds of inferences needed for effective word sense disambiguation.

The ideal situation for network-based methods is one where there are no intermediate nodes. That is, when each pair of appropriate word senses in a sentence is connected by a direct link in the network, and no word sense that is not to be selected (such as *stove)* has a direct link to any of the other selected senses. This, unfortunately, is rarely the case when processing real texts. As the number of intermediate nodes between desired senses of a pair of words increases, it becomes less likely that constraints represented by the intermediate nodes have an effect on the final selection. That is why in any real implementation on sizable networks and vocabularies, the standard network based methods turn out to be a mere game of luck. Often there are poorer alternatives that just happened to be cheaper. The only control is in adjusting weights on links which does not always guarantee that dynamic contexts will be inferred correctly to yield good results for previously unseen inputs.

5. Inferring and Applying Dynamic Contexts

Our method of checking selectional constraints exhaustively examines all the pairwise constraints on all word senses in a sentence, encoded statically in the network or in the lexicon, using a very efficient search mechanism, called Hunter-Gatherer, based on constraint satisfaction, branch and bound, and solution synthesis methods (Beale, Nirenburg and Mahesh, 1996). To augment this method to infer dynamic contexts, we introduce the Context Specialization Operator (CSO) with the following content: If a sense *P* is selected for a word *w,* and the rest of the word senses in the environment satisfy the constraints on *P*, examine the constraints on children of *P;* if exactly one child *C* of *P* satisfies the constraints, then infer that the correct sense of the word is *C,* apply the constraints on *C* to other words.

The semantic analyzer checks selectional restrictions and applies the CSO iteratively, thereby resolving word sense ambiguities successively. For example, "cake" is first determined to be a kind of *baked-food;* then using this information, "prepared" is disambiguated to *prepare-food.* Applying the CSO at this point shows that the only child of *prepare-food* which satisfies the constraint that the theme must be a *baked-food* and the instrument some sense of "range" was *bake.* Hence *bake* is included in the dynamic context and its constraints are applied to "range" in turn, thereby excluding *stove* and selecting *oven.*

6. Richer Knowledge: Current Methods Fizzle Out

As an example of interference from irrelevant knowledge, consider a situation where there is a direct link (such as *manufactured-by)* from stove to *human* in Figure 1. A similar link from *oven* to *human* may be missing in the knowledge base. Such omissions and inaccuracies are inevitable in any manually acquired knowledge base of nontrivial size. Marker passing and spreading activation will be unduly influenced by this additional link to *stove* and prefer *stove* as the meaning of "range" even though the additional link is completely irrelevant to the input at hand. It would be relevant if the input referred to a manufacturing relationship between "John" and "range." High connectivity in world knowledge representations is essential for general purpose NLP. Presumably, a link between any two nodes in the representation may at one time or another serve as a useful selectional constraint for word sense disambiguation.

Methods used to process such world knowledge must remain efficient even as network connectivity grows. Several attempts have, in fact, been made to contain the combinatorial effects in large-sized networks by constraining marker passing (e.g., Yu and Simmons, 1990). While such methods make processing more tractable, they do not help maintain its effectiveness at higher connectivities. Our method is robust against irrelevant connections in the knowledge base. It only considers those constraints that are relevant to the current text.

7. Implementation and Discussion

The methods outlined above have been implemented in a large-scale semantic analysis system in the Mikrokosmos machine translation project (Beale, Nirenburg and Mahesh, 1995; Mahesh et al, submitted). The system employs an ontological world model represented as a network of 5,000 concepts where each node has an average connectivity of 16 (Mahesh and Nirenburg, 1995). A Spanish lexicon of about 37,000 word senses maps to nodes in this network. The methods are not only very effective in resolving word sense ambiguities but are also very efficient. The system has been tested successfully on several real-life Spanish texts (each about 350 words long) in the domain of business news. A typical text had about 50 ambiguous (open-class) words of which roughly 97% were disambiguated correctly despite the fact that our model does not yet include reference resolution or discourse processing components.

One might argue that a simpler solution for our featured example would have been to edit the network and add a direct link between *oven* and *baked-food.* It is certainly possible to fine-tune the network or tweak the weights on the links to obtain a selection of oven. However, such an approach does not guarantee that desired results will be obtained outside training corpora. Moreover, such tuning invariably has a catastrophic effect on processing other inputs. For example, if we fixed the network so that *oven* is closer to *baked-food* than *stove,* then *oven* would be selected even in an example such as "John ate the cake on the range." There is, in fact, no information in this sentence that leads to a preference for either the *stove* or the oven sense of "range."

Statistical methods based on sense-tagged corpus analysis (e.g., Yarowsky, 1992) also appear to suffer from the same drawbacks as network search methods. In a sufficiently general corpus, collocations of word senses may lead to irrelevant interference in sense disambiguation. For example, a high degree of collocation between *baked-food* and *oven* helps select the right sense of "range" in "John prepared a cake with the range." However, the same statistical preference can mislead the processor into selecting the *oven* sense of *range* in "John ate the cake on the range," just as fine-tuning networks did.

Inferences of the kind proposed here can be recorded as "canned" episodic or stereotypical knowledge structures such as scripts, plans, and MOPs (e.g., Schank and Abelson, 1977). However, this strategy may not be easily scalable to sufficiently broad domains. Acquisition of script-like structures is often prohibitively expensive. The kind of connections shown in Figure 1, on the other hand, can be acquired at a reasonable cost, as the experience of the Mikrokosmos project shows (Mahesh and Nirenburg, 1995).

Note that our method would still benefit from scripts, if they are made available; not least for solving problems in discourse and pragmatics. The inference operator has potential applications in machine translation as well. If an instance of a relatively more specific concept, such as *bake* (relative to *prepare-food),* is included in the meaning representation of the source text which serves as interlingua text in an MT system, a language generator can use this clue to generate a more appropriate word in the target language than if *prepare-food* is included (for example, when the target language has different words for baking and other forms of cooking). We are currently investigating other possibilities of dynamic combination and modification of selectional restrictions to enhance their effectiveness in ambiguity resolution.

The above analysis assumed having a node for a lower concept (such as *bake)* to anchor the specific constraints on it. We are also investigating ways of representing such inter-argument constraints when such a node is absent, for instance, by automatically acquiring such nodes to bootstrap a knowledge-based NLP system and tune its knowledge base to a specific corpus.

8. Conclusion

Years of effort have made it possible to build large-scale knowledge bases for applications in NLP. However, current NLP methods were not designed with such knowledge in mind and are largely ineffective in solving difficult problems such as word sense disambiguation. In this article, we presented a way of using rich knowledge of semantic cooccurrence constraints in word sense disambiguation. This method effectively applies available knowledge by inferring dynamic contexts to resolve word sense ambiguities. We have also found that this method can be applied efficiently, using the Hunter-Gatherer control architecture. We believe that this line of work can lead to elegant models of solving semantic problems in NLP without placing unreasonable demands on knowledge acquisition.

References

Beale, S., Nirenburg, S. and Mahesh, K. (1996). Hunter-Gatherer: Three Search Techniques Integrated for Natural Language Semantics. In *Proc. Thirteenth National Conference on Artificial Intelligence* (AAAI96), Portland, Oregon.

Beale, S., Nirenburg, S., and Mahesh, K. (1995). Semantic Analysis in the Mikrokosmos Machine Translation Project. In *Proc. of the 2nd Symposium on Natural Language Processing,* 297-307. Bangkok, Thailand.

Charniak, E. (1983). Passing Markers: A Theory of Contextual Influence in Language Comprehension. *Cognitive Science* 7:171-190.

Charniak, E. (1986). A neat theory of marker passing.In *Proc. 5th National Conf. on Artificial Intelligence.* pp 584-588.

Eiselt, K. P. (1989). Inference Processing and Error Recovery in Sentence Understanding. Ph.D. Thesis, Computer Science Department, Univ. of California, Irvine, Tech. Report. 89-24.

Guthrie, L., Pustejovsky, J., Wilks, Y., and Slator, B. M. (1996). The role of lexicons in natural language processing. *Communications of the ACM,* 39(1): 63-72.

Knight, K. and Luk, S. K. (1994). Building a Large-Scale Knowledge Base for Machine Translation. In *Proc. Twelfth National Conf. on Artificial Intelligence,* (AAAI-94).

Mahesh, K. (1996). *Ontology Development: Ideology and Methodology.* Technical Report MCCS-96-292, Computing Research Laboratory, New Mexico State University.

Mahesh, K. and Nirenburg, S. (1995). A situated ontology for practical NLP. In *Proceedings of the Workshop on Basic Ontological Issues in Knowledge Sharing, International Joint Conference on Artificial Intelligence* (IJCAI-95), Montreal, Canada, August 1995.

Mahesh, K., Nirenburg, S., Beale, S., Viegas, E., Raskin, V., and Onyshkevych, B. (Submitted). Word Sense Disambiguation: Why Statistics When We Have These Numbers? Submitted to TMI-97.

Onyshkevych, B. and Nirenburg, S. (1995). A Lexicon for Knowledge-based MT. *Machine Translation,* 10: 1-2.

Schank, R. C. and Abelson, R. P. (1977). *Scripts, plans, goals, and understanding.* Hillsdale, NJ: Lawrence Erlbaum.

Waltz, D. L. and Pollack, J. B. (1985). Massively parallel parsing: A strongly interactive model of natural language interpretation. *Cognitive Science,* 9:51-74.

Yarowsky, D. (1992). Word sense disambiguation using statistical models of Roget's categories trained on large corpora. In *Proc. 14th International Conference on Computational Linguistics,* pp. 454-460, Nantes, France.

Yu, Y. and Simmons, R. F. (1990). Truly Parallel Understanding of Text. In *Proc. Eighth National Conference on Artificial Intelligence,* pp. 996-1001, Boston, MA, 1990. AAAI Press/The MIT Press.