Constructing Bayesian Networks from WordNet for Word-Sense Disambiguation: Representational and Processing Issues *

Janyce Wiebe and **Tom O'Hara**
Department of Computer Science and
Computing Research Laboratory
New Mexico State University
Las Cruces, NM 88003
wiebe, tomohara@cs.nmsu.edu

Rebecca Bruce
Department of Computer Science
University of North Carolina at Asheville
Asheville, NC 28804-3299
bruce@cs.unca.edu

Abstract

This paper describes a probabilistic model that is formed from the integration of an analytical and empirical component. The analytical component is a Bayesian network derived from WordNet, and the empirical component is composed of compatible probabilistic models formulated from tagged training data. The components are integrated in a formal, uniform framework based on the semantics of causal dependence. The paper explores various representational issues that must be addressed when formulating a Bayesian network representation of lexical information such as that expressed in WordNet. These issues are essential to the design of such a network and they have not been previously explored. We describe two choices for the representation of lexical items and two choices for the representation lexical relations. The effect of each combination of choices on evidence propagation in the network is discussed.

1 Introduction

There is a long tradition in AI of resolving interdependent lexical ambiguities through spreading activation, from Quillian's (1968) seminal work on semantic networks, through Hirst's work (1988) on Polaroid words, to more recent work by Voorhees (1993) and Veronis and Ide (1990) on large-scale disambiguation. This research investigates a probabilistic realization of spreading activation to resolve interdependent word-sense ambiguities. The core idea is to exploit belief propagation in Bayesian networks: Words are mapped to nodes, lexical relations are mapped to edges, and evidence is propagated from word senses to other related word senses.

The lexical relations are derived from an existing knowledge source, because this information cannot be automatically extracted from training data with existing techniques. The knowledge source we use is the WordNet *is-a* hierarchy, i.e., the *hypernym/hyponym* taxonomy (Miller, 1990). Although this hierarchy was developed for other purposes, it has been frequently applied to word-sense disambiguation (Resnik, 1995; Sussna, 1993). In this work, we investigate various approaches to constructing a Bayesian network representation of the *is-a* hierarchy for use in word-sense disambiguation. As this work continues, other relations such as part/whole and entailment relations will also be included in the network.

Another contribution of our work is a novel proposal for integrating symbolic and statistical information for the purpose of performing NLP tasks. Statistical approaches to word-sense disambiguation have had the most success to date, when evaluated on unseen test data. The "analytical" Bayesian network component of our method is actually built on top of "empirical" probabilistic classifiers induced statistically from training data. In particular, an empirical classifier is induced for each word in the current sentence to be disambiguated (i.e., for each *target word*). Each empirical classifier is developed independently of the empirical classifiers for other target words. A Bayesian network is constructed from the segment of the WordNet *is-a* hierarchy that is connected to the target words. The results of the empirical classifiers are fed as evidence into the Bayesian network, thus initiating belief propagation. All of the information is represented in a formal, uniform framework: a probabilistic model embodying conditional independence relationships among the variables that form the joint distribution. Conditional independence relationships simplify the formulation of the joint distribution making it possible to work with a large number of variables. Further, models that characterize conditional independence relationships have desirable computational properties (e.g., see the discussion on decomposable models in (Pearl, 1988)). These properties form the basis of the evidence propagation scheme used for Bayesian networks discussed in Section 7. We also make use of these properties in formulating the em-

* In *Proc. COLING-ACL '98 Workshop on the Usage of WordNet in Natural Language Processing Systems*, Association for Computational Linguistics, Montreal, Canada, August 16, 1998. This research was supported in part by the Office of Naval Research under grant number N00014-95-1-0776.

pirical classifiers as described in (Bruce and Wiebe, 1994). Bayesian networks are a very rich and complex representational framework. They support easy integration of diverse information sources and form the basis for much of the current work on reasoning under uncertainty (Pearl, 1988).

This paper explores the representational issues that must be addressed when mapping the lexical information in WordNet to a Bayesian network. The implications of the various choices are analyzed in depth. In section 2, we introduce the basic concepts and illustrate them with an example in section 3, which also includes a brief description of the empirical component. The Bayesian network representations of lexical items and lexical relations are discussed in sections 4 and 5, respectively. In section 6, we describe the integration of the empirical component into the Bayesian network The process of sense disambiguation is described in section 7. Section 8 discusses related work followed by conclusions in section 9.

2 Bayesian Networks: Background

Bayesian networks model dependencies among nodes through the use of conditional probabilities. Specifically, if a node ($Cause2$) is considered as a cause for another node ($Symptom1$), then the second node is defined relative to the first (i.e., $P(Symptom1|Cause2)$). Some nodes don't have associated causes, so they are just defined via unconditional probabilities (e.g., $P(Cause2)$). Taken together, the set of all the conditional and unconditional probabilities determine a joint distribution for all the nodes being modeled (e.g., $P(Symptom1, ..., SymptomN, Cause1, ...CauseM)$). Such global distributions are usually difficult to assess directly; hence, the Bayesian network provides a convenient formalism for specifying the same distribution via local distributions, under conditional independence assumptions. Furthermore. without the conditional independence relations, the full joint distribution for cases with hundreds of senses would be infeasible to process—the independence assumptions are key. Pearl (1988) presents an in-depth coverage of the theory of Bayesian networks and provides an efficient algorithm for evaluating them.

In a Bayesian approach to statistical inference, we distinguish between *prior* and *posterior* probabilities. Prior probabilities express the beliefs that we hold about the likelihood of events *prior* to being given any evidence, *posterior* probabilities express our beliefs in the likelihood of events given all the evidence that is currently known. Thus, the posterior probability of an event changes as new evidence is learned. The conditional and unconditional probabilities mentioned above are the prior probabilities. The posterior probabilities are calculated using the Bayesian network propagation algorithm each time new evidence is added. We discuss propagation in greater detail in Section 7. Intuitively, the posterior probability of a node, say the node GATHERING#1 (switching to a word-sense disambiguation example), is a combination of the beliefs received from its children and the beliefs received from its parents. Once a node has calculated its own belief, it calculates outgoing messages to send to its parents and to its children, which enable them, in turn, to calculate their posterior probabilities. In this way information is propagated throughout the network.

3 An Example

In this section, we illustrate how a simple Bayesian network can be constructed to model the interdependencies among words. This identifies the basic steps in the overall process and helps to motivate the representational issues discussed later.

Suppose that the words "community" and "town" appear in a single sentence, and that their correct senses in that context are COMMUNITY#1 and TOWN#2, respectively. Our task is to assign the correct word senses to both of them, considering information automatically derived from the corpus and gathered individually for each word, as well as information derived from the WordNet *is-a* hierarchy and represented in a Bayesian network. The basic strategy is to add the corpus-derived information to the Bayesian network representations of "community" and "town," in such a way that it initiates propagation.

Let us consider this process in more detail. The words "community" and "town" have the following senses in WordNet:

community:
1. people living in a particular local area
2. an association of people with similar interests
3. common ownership
4. the body of people in a learned occupation

town:
1. an urban area with a fixed boundary that is smaller than a city
2. the people living in a municipality smaller than a city
3. an administrative division of a county

These senses are represented as sets of synonyms, or *synsets*. In the *is-a* hierarchy, each synset is linked to its *hypernym*, i.e., the synset representing its conceptual parent. For example, the synset corresponding to

{occupation, vocation, occupational group}

is the hypernym of the synset corresponding to

{profession, community}.

A new Bayesian network is created for each sentence. It includes all of the synsets for the target words in the sentence, together with all of the synsets reachable from them in the WordNet *is-a* hierarchy. Extracting this information from WordNet is straightforward.

Figure 1 illustrates one way that the Bayesian network for the example sentence containing "town" and "community" can be constructed. In this representation, each word sense is mapped to a node in the network, and there is an edge from X to Y iff word sense X is a hypernym (i.e., a superordinate) of word sense Y (please ignore the octagonal nodes at the bottom for now). Notice that the relation between COMMUNITY#1 and TOWN#2 is mediated by GATHERING#1, a type of GROUP#1. Our goal is for the contextual evidence provided by the empirical classifiers to propagate along this path in such a way that the correct senses of the target words reinforce one another.

After the topology of the network has been established, the conditional probability tables required for each node must be defined. As will be discussed later in section 5, we can make independence assumptions that make estimating the necessary probabilities more easier.

Next, an empirical classifier is developed for each ambiguous word, in this case, "town" and "community". Each classifier defines a probability distribution describing the likelihood of each sense of the targeted word given the automatically derived features of the context. An example of the type of feature used is the part-of-speech of the word to the right; see (Bruce and Wiebe, 1994) for the other ones we use.

The distributions determined by the empirical classifiers are added as evidence to the Bayesian network, initiating belief propagation. Once the network reaches equilibrium, the posterior probabilities of the nodes for "town" and "community" determine the senses assigned to each ambiguous word.

4 Representing Lexical Items: What does a Node mean?

There are two basic approaches to representing WordNet synsets in a Bayesian network. Since the lexical relations are among synsets and not words, a natural approach is to represent the synsets as nodes. Alternatively, one node could be used to represent all senses of a word.

4.1 The One Node Per Word Approach

When nodes correspond to words, the possible values for each node are *sense*0 through *senseN*, where N is the number of WordNet synsets representing senses of the target word. *Sense*0 represents the composite of all other meanings, i.e., of all meanings that are not represented by WordNet synsets. Figure 2 shows the graph for the Bayesian network when word nodes are used for the relations. It also illustrates the use of logical links, which are described in the next section. This involves more than just a change in link direction.

4.2 The One Node Per Sense Approach

Figure 1 illustrates the approach in which each synset (each sense) is mapped to a node. An important advantage of using the node per sense approach is that it facilitates handling dependencies among the senses of a word. In the node per word approach, single node cycles are produced when modeling the dependencies of words that have a meaning that is defined in terms of other meanings for that same word.

A disadvantage of this approach is that modeling mutual exclusion among the senses of a single word becomes more difficult. The most straightforward approach modeling mutual exclusion is to create a dependency from each sense node to a separate node with a CPT enforcing mutual exclusion. But since the table must have 2^N entries, this approach becomes impractical for words with a large number of senses. To get around this problem, two levels of mutual-exclusion dependencies could be introduced: one at which mutual exclusion among small groups of senses is enforced, and another enforcing mutual exclusion of the groups.

5 Representing Lexical Relations: What does an edge mean?

Here, we address issues concerning the representation of WordNet *is-a* relationships as causal dependencies. The two primary issues to be addressed are: (1) expressing the Hypernym/Hyponym relationship as a causal dependency, and (2) quantifying the causal dependencies with conditional probability distributions.

5.1 Hypernym→Hyponym Representations

The Hypernym → Hyponym Representation was illustrated above in section 3: there is an edge from node X to node Y iff X represents a hypernym of node Y in the WordNet *is-a* hierarchy. Consider the node per sense representation (see figure 1). Suppose $Hyper$ is a synset that is a hypernym of synsets $Hypo_1 \cdots Hypo_k$. Then, the relevant part of the Bayesian network expresses the following:

$$Hyper \rightarrow Hypo_1 \vee \cdots \vee Hypo_k$$

As such, we are making a closed world assumption. If, for example, there is a synset ANIMAL#1 with three hyponyms DOG#1, CAT#1, and MOUSE#1,

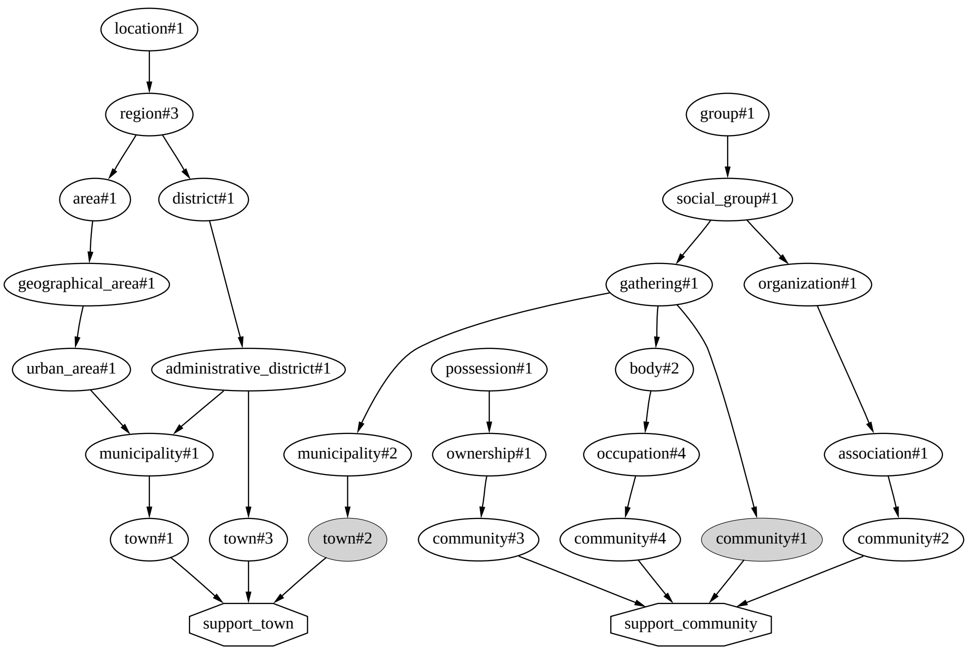

Figure 1: Sense per node Bayesian network with hypernym→hyponym links

we are assuming that these three are the only kinds of ANIMAL#1's there are.

When using this link representation with either of the node per sense or the node per word representations, the roots of the network are the most superordinate synsets reachable from the target words, and the target words are typically (but not necessarily) the leafs of the network.

We now turn to defining the CPT. We discuss this with respect to the node per sense representation (in figure 1) because it is easier to discuss and similar conditional probabilities must be defined under the node per word representation.

To define the CPT for each child node in the Bayesian network, where each child node corresponds to a hyponym node in WordNet, we assign the conditional probability $P(hyponym|hypernym)$ to be inversely proportional to the number of children that the hypernym has. For instance, MUNICIPALITY#1 has two children in WordNet, so we assign the following conditional probability for TOWN#1 given this hypernym.

P(town#1 | municipality#1)

municipality#1	P(town#1)
F	$0.000 + \epsilon$
T	0.500

In so doing we are: (1) considering each hyponym of a given hypernym to be equally likely, and (2) maintaining the closed world assumption by requiring that these conditional probabilities sum to one. In all CPTs, we add a small positive probability ϵ to all zero probability values in order to allow the realization of all possible configurations of node values (e.g., to handle inconsistent evidence). In future work, we will consider using frequency of occurrence information in tagged training data to define these CPTs.

For the root nodes, which represent the most superordinate concepts, prior probabilities must be specified. With no evidence to the contrary, uniform prior distributions are assigned to the root nodes; the empirical classifiers are relied upon to provide contextual support (through the leafs of the network).

5.2 Hyponym→Hypernym Representations

Under the Hyponym → Hypernym Representation, there is an edge from node X to node Y iff X represents a hyponym of node Y in the WordNet *is-a* hierarchy. Consider the node per sense representation (see figure 2). The Bayesian network represents the following:

$$(Hypo_i = s_i \rightarrow Hyper_j = s_j)$$
$$\wedge \cdots \wedge (Hypo_n = s_n \rightarrow Hyper_m = s_m)$$

Under the semantics of the WordNet *is-a* hierarchy, all instances of a hyponym are instances of its

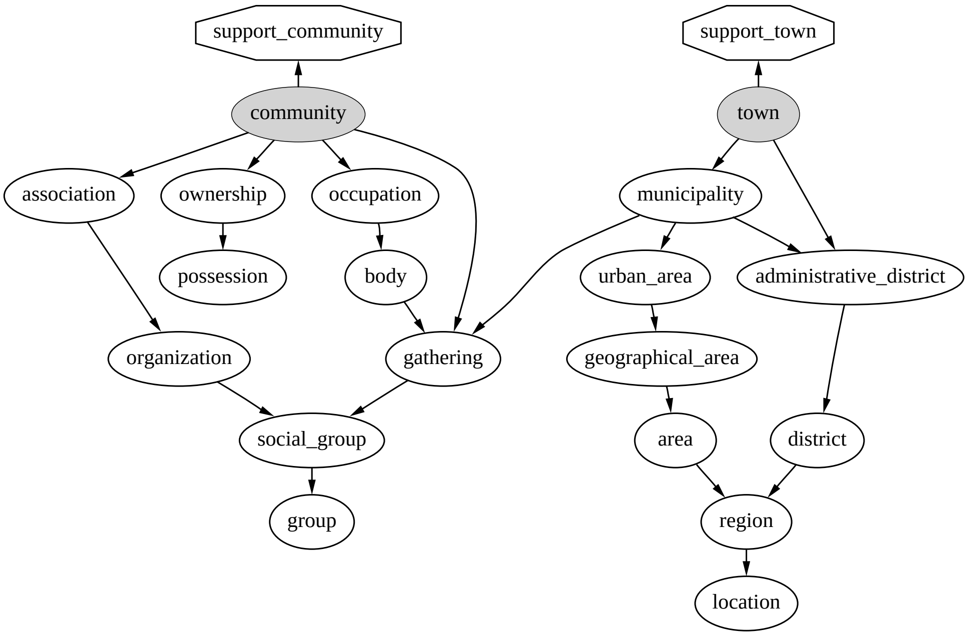

Figure 2: Word per node Bayesian network with hyponym→hypernym links

hypernym. So, a typical CPT for this representation is as follows:

P(municipality#1 | town#1)

town	P(municipality#1)
F	0.0 + ϵ
T	1.0 - ϵ

Note that this case is not illustrated in the graphs shown: these only cover two of the four main possibilities.

Interestingly, in this representation, the root nodes represent the target words. Thus, the root nodes are the sites where evidence from the empirical classifiers is added to then network. In the absence of this evidence, these nodes take on their prior probabilities. As above, we assign uniform distributions as the priors. Recall that, in the case of multiple parents, CPTs must specify the conditional distribution of the child node given the values of all of its parent nodes. The issues involved in working with multiple parent nodes are discussed below.

5.3 CPT Entries when Multiple Parents: Causal Independence

If a node has multiple parents, say n parents, then specifying all of the entries in the CPT for that node can be prohibitive. If no additional independence assumptions are made regarding the interactions among the parent nodes, then the number of probabilities that must be specified is exponential in n, and probabilistic inference is made correspondingly more complex (Heckerman and Breese, 1994). To overcome this problem, the *noisy-OR* model (Pearl, 1988) is often adopted. Under this model, certain independence assumptions are made regarding the interactions among the parent nodes, with the effect that the number of probabilities that must be specified is linear in n. Basically, one need only specify the conditional probabilities of the child and each parent individually.

As presented in (Pearl, 1988), the noisy-OR model assumes that all of the variables are binary. Heckerman and Breese (1994) present a generalization of the noisy-OR model, *causal independence*. In this model, the parents are assumed to be independent causes for the child. This allows us to formulate a CPT from the specification of only the following conditional probabilities: $P(c|p_{ij})$, where c ranges over the values of the child, and p_{ij} ranges over the values of parent P_i. These values are combined via the constraints of the model to produce the CPT for the child node.

We assign the probabilities using a causal independence model which specializes to the noisy-OR model when applied to binary nodes. First consider that the inclusive-or connective can be viewed as outputting a true value iff none of the inputs is false:

$$output = \neg((\neg v_1) \wedge \cdots \wedge (\neg v_n))$$

where each v_i is a logical-valued input variable. The extension to the case where probabilities are associated with each input is relatively straightforward:

$$child = \neg((\neg v_1) \wedge \cdots \wedge (\neg v_n))$$
$$P(child|V_1 = v_1, ..., V_n = v_n) =$$
$$1.0 - \Pi(1.0 - P(child|v_i)), \qquad \forall_{V_i} v_i = T.$$

When extending to the general case, the relationship between the value of the child node and the values of its parent nodes is not necessarily defined by a truth function. But, the probabilities are assigned analogously:

$$P(Child = c|V_1 = v_1, ..., V_n = v_n) =$$
$$1.0 - \Pi(1.0 - P(child = c|V_i = v_i))$$
$$\forall_{V_i} P(child = c|V_i = v_i) > \epsilon.$$

In their work on plan recognition, Charniak and Goldman (1993) use the noisy-OR model, specifically for representing the dependencies of observed actions on the potential plans that could explain them.

6 Integrating Empirical and Analytical Information: Virtual Evidence Nodes

Due to space limitations, we consider just one method for integrating the empirical and analytical components. In this technique, support from the empirical classifiers is added to the Bayesian network using *virtual evidence nodes* (Pearl, 1988). The usual way to add evidence to a Bayesian network is to instantiate a node to a particular value (called "clamping"); the influence of this evidence is then propagated through the network. However, that method is not appropriate for our task, because we do not know the sense of any word (so there is no node in the Bayesian network that can be initially instantiated). Virtual evidence nodes provide a way to specify uncertain evidence, in the form of a distribution over node values (i.e., the probability of each node value). They are represented by the octagonal nodes in figures 1 and 2. There is one for each of the target words to be disambiguated. These nodes represent the support for each sense that was derived from the corpus by the empirical component. Each virtual evidence node is implemented as a binary-valued node whose parent is the node for which evidence is being provided. The evidence distribution determines the conditional probability table.

7 Edge Direction and Belief Propagation

There is a very important implication of the choice between the hypernym → hyponym and the hyponym → hypernym representations. In a Bayesian network, suppose that evidence is added to a node (either by clamping or by virtual evidence nodes). This evidence will propagate to its ancestors in the Bayesian network, **and also to the children of its ancestors**. For example, in figure 1, evidence introduced at node SUPPORT_COMMUNITY will propagate, among other places, back to COMMUNITY#1, back to GATHERING#1, and then down to MUNICIPALITY#2, and so on. Thus, this representation, hypernym → hyponym, supports the kind of propagation described in this paper.

On the other hand, consider the hyponym → hypernym representations (figures 1 and 2). In these representations, the targeted words are the roots of the Bayesian network, so the evidence is added to the roots of the network. This evidence will **not** propagate from, say, COMMUNITY#1 to TOWN#2 in figure 1. Information propagates between such nodes **only if evidence were added to their mutual descendents**. As Pearl says, "evidence gathered at a particular node does not influence any of its spouses until their common child gathers diagnostic support" ((Pearl, 1988), p. 182). Thus, if evidence is only added at the virtual evidence nodes in figure 2, evidence will not propagate from COMMUNITY=1 to MUNICIPALITY=2 (so it will not propagate further to TOWN=2). The corresponding nodes are spouses, but their child (GATHERING) has not received diagnostic support, by which Pearl means evidence propagated from below.

However, there are many other possibilities for adding evidence to the network, under which desired propagation would occur. Thus, our discussion of the hyponym → hypernym representations is not just a cautionary tale. For example, one might use Yarowsky's (1992) unsupervised method for assigning words to thesaural categories to add evidence to a node representing a superordinate concept in the WordNet *is-a* hierarchy. (Virtual evidence nodes could be used for this purpose too.) In the hyponym → hypernym representations, this superordinate concept (say GATHERING#1 or SOCIAL-GROUP#1) is a **descendent** of the nodes representing the targeted words. It would thus provide the needed diagnostic support to enable propagation from one target word to another. Note that the hyponym → hypernym representation is conceptually appealing, since its semantics is based directly on the semantics of the WordNet *is-a* hierarchy.

As an illustration, consider applying sample evidence of (.70, .10, .10, .10) for the senses of "community" (with no evidence for town). Table 1 shows the posterior probabilities before and after applying this evidence.

As can be seen, the high evidence for COMMUNITY#1 increases the support for the hypernym GATHERING#1 (as well as for the other ancestors in the same path not shown). However, no support is reaching MUNICIPALITY#2.

If the hypernym → hyponym representation is used instead (as in figure 1), an appropriate propagation does take place. The propagation occurs in

sense	before	after
community#1	.20	**.70**
gathering#1	.55	.87
municipality#2	.25	.33
town#2	.25	**.33**
community#4	.20	.10
body#2	.20	.10
location#1	.50	.67
municipality#1	.25	.33
town#1	.25	**.33**

Table 1: Propagation w/ hyponym→hypernym links

sense	before	after
community#1	.054	**.562**
gathering#1	.126	.631
municipality#2	.063	.312
town#2	.060	**.290**
community#4	.020	.030
body#2	.064	.296
location#1	.500	.500
municipality#1	.068	.063
town#1	.053	**.050**

Table 2: Propagation w/ hypernym→hyponym links

two phases. First, the high evidence for COMMUNITY#1 is propagated "upstream" to the hypernym node. Then, the increased support for this synset is propagated "downstream" to increase the likelihood of the value for the appropriate sense of "town". Table 2 shows the posterior probabilities in this case.

7.1 Attenuation of Spreading Activation

An important aspect of spreading activation approaches is that the strength of the evidence being propagated is attenuated the further the evidence spreads from the original source. Traditional spreading activation schemes have used various heuristics to model this attenuation, often incorporating a distance factor in terms of number of links. By using probabilistic propagation, we can account for both length of path and fan-out at the nodes along the path (i.e., how many children they have). The length of the path is taken into account by the propagation algorithm. Intuitively, when a node calculates its posterior distribution, it calculates a distribution taking into account all possibilities (e.g., gathering#1=1, municipality#2=1; gathering#1=1, municipality#2=0; and so on). As the evidence is dispersed among the various possibilities at subsequent nodes, the evidence for any single possibility tends to decrease. This is so for either edge direction.

8 Comparison to Related Work

Spreading activation schemes have been common in various forms, starting with Quillian's (Quillian, 1968) work on semantic memory. Quillian used spreading activation to identify paths between concepts for the purpose of comparison and contrast. To construct the semantic networks, dictionary definitions were manually encoded in the form a graph.

Hirst (1988) also used spreading activation to perform word-sense disambiguation. The approach relies on the identification of paths between interdependent word meanings. To avoid extraneous connections, constraints were introduced; for instance, a limit on path length was introduced, and *is-a* links were normally not traversed in reverse direction. Furthermore, heuristics were used to give preference to shorter paths and to avoid connections through nodes with many out-going arcs.

There have been several approaches that have relied upon word-overlap in dictionary definitions to resolve word-sense ambiguities in context, starting with (Lesk, 1986). Cowie et al. (1992) extend the idea by using simulated annealing to optimize a configuration of word senses simultaneously in terms of degree of word overlap.

Veronis and Ide (1990) developed a neural network model to overcome the limitation of addressing only pairwise dependencies in word-overlap approaches. Using dictionary definitions, they constructed a network containing links from each word node to the nodes for each of its senses and links from each of the sense nodes to the nodes of the words used in the definition.

Sussna (1993) produces a semantic network based on several different WordNet relations. His disambiguation method minimizes the pairwise distance among senses via a weighting scheme that accounts for both fan-out and depth in the hierarchy. Of the approaches we have surveyed, his is most similar to our analytical component.

Voorhees (1993) describes an unsupervised approach that exploits the WordNet hypernym taxonomy. In particular, the hierarchy for a given word is automatically partitioned so that the words occurring in the synsets of a partition (or *hood*) only occur with one of the senses for the word. Disambiguation is based on the selecting the hood which has the highest estimated relative frequency for the context relative to training text.

Resnik (1995) also describes an unsupervised approach that is based on estimating synset frequencies. As with Voorhees, the estimated frequency of a synset is based on the frequency of the word plus the frequencies of all its descendant synsets in a large corpus. Therefore, the top-level synsets have the highest frequencies and thus the highest estimated frequency of occurrence. For each pair of nouns from the text to be disambiguated, the *most-informative-subsumer* is determined by finding the common ancestor with the highest information content, where information content is inversely related to frequency. Then each noun is disambiguated by selecting the synset that receives the most support (i.e., information content) from the all of the most-informative-subsumers.

Eizirik et al. (1993) also describe a Bayesian network model for word-sense disambiguation, which includes syntactic disambiguation as well as lexical information. However, their networks are not automatically constructed.

9 Conclusion

This paper explores various representational issues that must be addressed when formulating a Bayesian network representation of lexical information such as is expressed in WordNet. We describe two choices for the representation of lexical items and two choices for the representation lexical relations. The effects on evidence propagation in the network is also discussed.

References

Bruce, R., and Wiebe, J. (1994), "Word-sense disambiguation using decomposable models", in *Proc. of the 32nd Annual Meeting of the Association for Computational Linguistics (ACL-94)*, pp 139–146.

Charniak, E., and R. Goldman (1993), "A Bayesian Model of Plan Recognition", *Artificial Intelligence* 64:53-79.

Cowie, J., J. Guthrie, and L. Guthrie (1992), "Lexical Disambiguation Using Simulated Annealing", *Proc. COLING-92*, pp. 359-365.

Eizirik, L. , V. Barbosa, and S. Mendes (1993), "A Bayesian-Network Approach to Lexical Disambiguation", *Cognitive Science*, 17:257-283.

Heckerman, D. and J. Breese (1994), "Causal Independence for Probability Assessment and Inference Using Bayesian Networks", Technical Report MSR-TR-94-08, Microsoft Research, (Revised October, 1995).

Hirst, G. (1988), "Resolving Lexical Ambiguity Computationally with Spreading Activation and Polaroid Words", in *Lexical Ambiguity Resolution*, S. Small, G. Cottrell, and M. Tanenhaus (eds), San Mateo, CA: Morgan Kaufmann Publishers, pp. 73-107.

Lesk, M. (1986), "Automatic Sense Disambiguation Using Machine Readable Dictionaries: How to Tell a Pine Cone from an Ice Cream Cone", in Proc. SIGDOC, Toronto.

Miller, G., (1990), "WordNet: An On-line Lexical Database", *International Journal of Lexicography* 3(4).

Pearl, J. (1988), *Probabilistic Reasoning in Intelligent Systems*, San Mateo, CA: Morgan Kaufmann.

Quillian, M.(1968), "Semantic Memory", in *Semantic Information and Processing*, M. Minsky, ed., Cambridge, MA: MIT Press.

Resnik, P. (1995), "Disambiguating Noun Groupings with Respect to WordNet Senses", in *Proc. Third Workshop on Very Large Corpora*, Cambridge, MA, June 1995.

Sussna, M. (1993), "Word Sense Disambiguation for Free-text Indexing Using a Massive Semantic Network", in *Proc. Second International Conference on Information and Knowledge Management (CIKM-93)*, Arlington, Virginia.

Veronis, J., and N. Ide (1990), "Word Sense Disambiguation with Very Large Neural Networks Extracted from Machine Readable Dictionaries", in *Proc. COLING-90*, Helsinki, August 1990.

Voorhees, E. (1993) "Using WordNet to Disambiguate Word Senses for Text Retrieval", in *Proc. 16th Annual ACM SIGIR Conference on Research and Development in Information Retrieval*, Pittsburgh, pp. 171-180.

Yarowsky, D. (1992), "Word-Sense Disambiguation Using Statistical Models of Roget's Categories Trained on Large Corpora", in *Proc. COLING-92*, Nantes, Aug 23-28, pp. 454-460.