ARTICLE IN PRESS

ELSEVIER

Cognitive Science 87 (2002) 1–13

COGNITIVE SCIENCE

http://www.elsevier.com/locate/cogsci

Statistical models for the induction and use of selectional preferences

Marc Light*, Warren Greiff

The MITRE Corporation, 202 Burlington Road, Bedford, MA 01730, USA

Accepted 22 March 2002

Abstract

Selectional preferences have a long history in both generative and computational linguistics. However, since the publication of Resnik's dissertation in 1993, a new approach has surfaced in the computational linguistics community. This new line of research combines knowledge represented in a pre-defined semantic class hierarchy with statistical tools including information theory, statistical modeling, and Bayesian inference. These tools are used to learn selectional preferences from examples in a corpus. Instead of simple sets of semantic classes, selectional preferences are viewed as probability distributions over various entities. We survey research that extends Resnik's initial work, discuss the strengths and weaknesses of each approach, and show how they together form a cohesive line of research. © 2002 Published by Cognitive Science Society, Inc.

Keywords: Computational linguistics; Selectional preferences; Statistical modeling; Learning

1. Introduction

Words in the same sentence stand in relationships with one another. For example, in *the person quickly ate the delicious sandwich*, the verbal predicate *eat* has *person* and *sandwich* as arguments. Similarly, *quickly* and *delicious* have as arguments *eat* and *sandwich*, respectively. These predicates have preferences for the semantic class membership of the arguments filling a particular role. For example, *eat* prefers, as its object argument, words from the semantic class of FOOD and disprefers words from the semantic class of FLUIDS.

* Corresponding author. Tel.: +1-781-271-5579; fax: +1-781-271-2352.
E-mail address: light@mitre.org (M. Light).

PII: S0364-0213(02)00070-8

In some sense, "selectional preferences" also exist in the other direction: arguments select for predicates. *Cake* prefers to be *baked* and not *written* in contrast to *books*. But most of the literature on selectional preference induction focuses on the preference of predicates for their arguments,[1] and the present literature review will do the same. For expository reasons we will further restrict our focus to the selectional preferences of transitive verbs for their object noun phrase argument.

Another restriction on the scope of this article is that we will assume that the semantic classes are given: they represent *pre-existing* world and lexical knowledge (see Fig. 1 for examples of semantic class membership and class subsumption knowledge). Thus, the work described here discusses how classes, possibly generated by other cognitive processes, can be used in language processing. In contrast, research such as Lee, Pereira, and Tishby (1993) discusses how semantic classes might be bootstrapped from language input.

The general idea of selectional preferences has been part of generative linguistics from the beginning (Katz & Fodor, 1964; Chomsky, 1965). It also has a long history in computational linguistics (Grishman, Hirschman, & Chomsky, 1965). However, since the publication of Resnik's dissertation (1993), a new approach has emerged in the computational linguistics community. This new line of research combines knowledge represented in a pre-defined semantic class hierarchy with statistical tools including information theory, statistical modeling, and Bayesian inference. Thus, *eat*'s preferred objects are represented not as the black-and-white class FOOD but rather as a gray probability distribution over all nouns or various classes thereof (or equivalently, as a stochastic model that generates some objects more often than others). Such definitions then suggest methods for learning selectional preferences from examples. These acquisition methods are computationally feasible, produce intuitively reasonable and demonstrably useful preferences, and can benefit from large amounts of possibly noisy data.

The availability of a large semantic hierarchy, WordNet (Fellbaum, 1998; Miller, 1990), made this work possible. WordNet is a thesaurus-like object that has classes that can be regarded, extensionally, as sets of words, and, intensionally, as elements in an abstract ontology. It has over 60,000 semantic classes with over 90,000 English words assigned to one or more classes. This is information that a human English speaker might be expected to have.

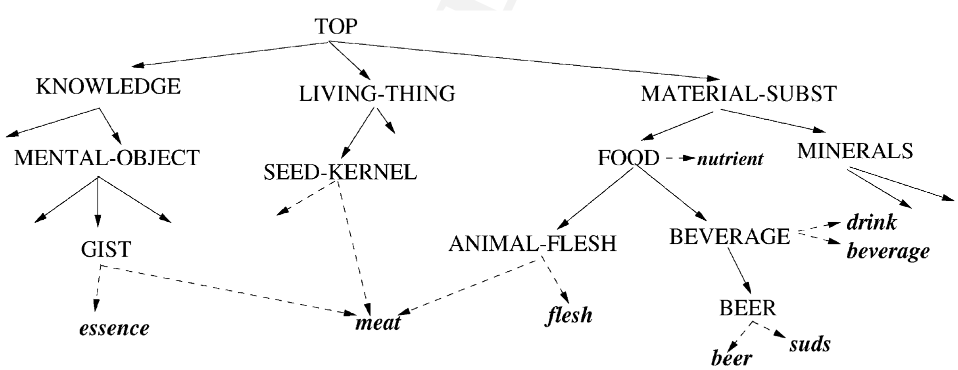

Fig. 1. An example semantic class hierarchy.

Table 1
Objects of *eat* in the BNC

Food	77	Bag	2	Investment	1
Meat	45	Dish	2	Kitchen	1
Meal	46	Hole	2	Mustard	1
Breakfast	30	Ice	2	Pack	1
Egg	18	Majority	2	Pasta	1
Bread	14	Proportion	2	Principle	1
Sandwich	13	Salad	2	Salt	1
Dinner	11	Scrap	2	Sauce	1
Slice	7	Soup	2	Sheep	1
Spaghetti	6	Trout	2	Stick	1
Chicken	5	Average	1	Sugar	1
Fry	4	Bucket	1	Tape	1
Roll	4	Feast	1	Top	1
Root	4	Fry	1	Yogurt	1
Mouthful	3	Garlic	1		

Equally important to the work described here was the availability of training material for the induction of the statistical models provided by large machine-readable corpora and tools for extracting verb argument pairs. As an example, statistics for objects of the verb *eat* are given in Table 1. Shown are objects of *eat* found in the British National Corpus (100 million words) (Burnard, 1995), together with their frequency of occurrence. These data were extracted using an automated partial parser (Abney, 1997).

This paper provides a survey of this line of research. We will look at Resnik (1993), Li and Abe (1998), Clark and Weir (1999), Abney and Light (1999), and Ciaramita and Johnson (2000). We hope to provide the newcomer an introduction and provide the expert an interesting juxtaposition of perspectives and methods used. Since the work originates from the field of computational linguistics, it often leaves unexplored ramifications for human language processing and acquisition.

Two central questions for the automated treatment of selectional preferences are: what *representation* to use, and how to *induce* preferences from available data. The representation of the selectional preferences can be thought of as a mapping, $\sigma : (v, r, c) \mapsto a$, that maps each selectional tuple (v, r, c) to a real number a; the degree of preference of a verb v for a class c with respect to role r. Examples are given in Table 2. Issues concerning *representation* include:

Table 2
Example selectional tuples

Predicate	Role	Semantic class	Weight
Eat	*Subj*	CAUSAL-AGENT	0.8
Eat	*On*	SURFACE	0.6
Eat	*Obj*	FOOD	0.9
Eat	*Obj*	BEVERAGE	−1

- What is the range of the weights a? For example, the range might be limited to the set $\{1, 0\}$ in which case the preferences are Boolean (black-and-white rather than gray).
- Where do the weights come from? For example, weights might be the parameters of a statistical model, estimated from the data.
- What is the interpretation of the representation? For example, weights may relate directly to the expected frequency of words appearing in the role.

Induction can be understood as how to use available data to decide what weight each (v, r, c) triple should receive. For example, if these weights come from a statistical model, then the induction process is equivalent to using the data to select a model and estimate its parameters.

A central problem for induction is noise in the training data: problematic examples that could lead induction astray. Noise can be due to errors in part of speech tagging or syntactic analysis, or due to metaphorical usage. Examples from Table 1 include the entries for *investment*, *average*, *tape*, and *race*. Typically, however, "good" examples such as *food* and *meal* will appear with much greater frequency.

Another central problem is word sense ambiguity in the training data. The word *bread* in Table 1 provides an example. *Bread* can be used to refer to a FOOD, e.g., *the multi-grain bread in Germany is wonderful*, but it can also refer to MONEY, e.g., *I could really use some bread since my car just broke down*. For this reason, it is not immediately clear whether the 14 tokens of *bread* in Table 1 provide evidence that *eat* subcategorizes for FOOD or for MONEY. If the wrong choice is made for a high frequency word, incorrect generalizations may result. Because the word sense for each token is not observable, the problem of inducing selectional preferences is said to involve incomplete data.[2]

We have discussed representation and induction but have not yet mentioned how selectional preferences fit into a larger picture of language processing. They are not an end in themselves but are a knowledge source for performing other language processing tasks. We give three examples below.

- *Syntactic structure*: the attachment of prepositional phrases is influenced by the selectional preferences of the heads of the attachment sites. For example, in *he bought the pants from the rack*, the attachment of the phrase headed by *from* could be based on the dispreference of ⟨*buy*, *from*⟩ for *rack*. *He bought the pants from the store* illustrates the alternate attachment.
- *Speech recognition*: in automatic recognition, the analysis of the acoustic signal is balanced against information about the likelihood of the sequence of words and the overall probability is maximized. Selectional preferences can influence how likely a sequence is. For example, given that *they ate* has been recognized, selectional preferences would make *peaches* more likely than *beaches* despite their acoustic similarity.
- *Word sense disambiguation*: words often have multiple meanings but for any given context, the choice is usually clear. Selectional preferences are part of the disambiguating context. For example, *meat* in *they ate the meat* refers to the ANIMAL-FLESH meaning (a subcategory of FOOD in Fig. 1) and not the GIST (e.g., *the meat of the argument*) meaning.

In general, selectional preferences allow semantic information to be used by other language processing components without requiring knowledge of the full complexity of the semantics of the lexical items and the interpretation of the surrounding utterance and dialogue. It seems plausible that successful experiments relevant to human language acquisition and processing could be carried out that are based on the work described here. Again Resnik has performed some initial work. In Resnik (1996), he demonstrated the following correlation: a transitive verb's strength of selection with respect to its object argument predicts how likely it is that this verb can also be used intransitively. For example, *eat* has a strong preference for foods as objects in comparison to the verb *make* which does not prefer any sort of object very strongly. Correspondingly, *John ate* is felicitous whereas *John made* is not. However, the work described here does not further address the ramifications for human language processing.

2. Approaches to inducing selectional preferences

The approaches described here represent a cohesive line of research. Resnik (1993) made use of WordNet (Miller, 1990), trained on corpora derived from the UPenn TreeBank parses of the Brown Corpus (Marcus, Santorini, & Marcinkiewicz, 1993). Furthermore, he used information theory to describe selectional preferences. Although, the use of probability distributions are central to Resnik's approach, there is no explicit statistical model for selectional preferences. In contrast, the remaining four papers do give explicit statistical models. Li and Abe (1998) use the minimal description length principle to pick a model that balances generality and accuracy with respect to the training data. Their work is also fully grounded in information theory. To the same end, Clark and Weir (1999) use statistical significance measures. The statistical models used by Abney and Light (1999) are hidden Markov models (HMMs). These HMMs are the first models to explicitly produce distributions over words as selectional preferences. From these, distributions over classes can be computed as well. In addition, they also deal with word sense ambiguity in the training data using an expectation maximization (EM) algorithm. Finally, Ciaramita and Johnson (2000) frame the problem as a Bayesian network and also deal with ambiguity in the training data.

2.1. Probability distributions, Kullback–Leibler divergence, and selectional association

Resnik (1993) initiated a new line of research explicitly concerned with induction of selectional preferences from training data and a class hierarchy such as WordNet. The result of his induction algorithm is the assignment of real numbers to the nodes of the hierarchy, indicating the degree of *selectional association* that classes have with respect to the verb.

The induction method makes use of two probability distributions over classes: $p(C)$ and $p(C|v)$. For each class c, the conditional probability $p(c|v)$ indicates how often a token of verb v takes a direct object in class c, whereas the marginal probability $p(c)$ indicates how often direct objects fall in class c in general. Selectional association weights are derived from these probability distributions. The intuition is that selectional association is greatest where the difference between the two distributions is largest: $p(c|v) \gg p(c)$ for a positive association, and $p(c|v) \ll p(c)$ for a negative one. For example, the probability of FOOD may be relatively small in the corpus in general, but jumps up considerably when looking only at nouns that are the object of *eat* (see Fig. 2).

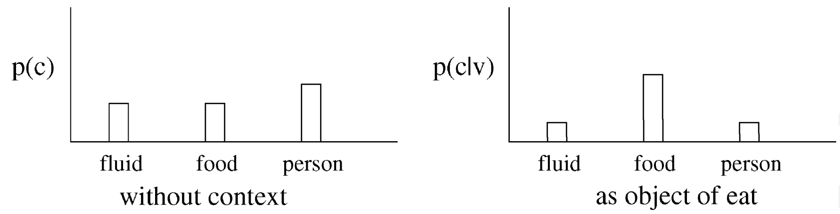

Fig. 2. Distributional changes (adapted from Resnik, 1997).

A measure borrowed from information theory, the Kullback–Leibler divergence,

$$D[p(C|v)||p(C)] \stackrel{\text{def}}{=} \sum_{c \text{ in Classes}} p(c|v) \log\left(\frac{p(c|v)}{p(c)}\right)$$

is used to measure the difference between the two distributions over classes. This aggregate difference is considered the *selectional preference strength* of the verb v. The *selectional association* of v for a specific class, c, is the contribution of that concept to the total selectional preference strength:

$$\mathrm{SA}(v, c) = \frac{p(c|v)\log(p(c|v)/p(c))}{D[p(C|v)||p(C)]}$$

It is the difference in the distributions at a particular class normalized by the sum of differences over all classes.

The estimation of the probability distributions may appear straightforward. For each class, c, $p(c|v)$ is estimated as $f(v, c)/\sum_{c'} f(v, c')$, where $f(v, c)$ is the number of times that verb v appears with a direct object in class c. Unfortunately, difficulties arise due to the word sense ambiguity in the data. The number of times a word of concept, c, occurred is not known because the appropriate sense is not indicated for ambiguous words. To address this problem, the counts for ambiguous words are divided equally among the possible classes for the word. For example, if *meat* is found to occur as the object of *eat* and is a potential member of nine classes, then a ninth of the total count is attributed to each class.[3] (Fractional counts may occur but are natural in a probabilistic framework.) Such a uniform allotment is an initial attempt to model uncertainty and turns out to produce reasonable results.

In sum, Resnik is the first to explicitly attack the problem of induction of selectional preference using a pre-existing semantic class hierarchy. Although a probabilistic approach is adopted, using a measure borrowed from information theory, induction cannot be said to result in the production of a statistical model that predicts the future objects of *eat*, as it does in the later efforts discussed below. In addition, word sense ambiguity in the training data is treated in an overly simple manner.

2.2. Statistical modeling, information theory, and hypothesis testing

Li and Abe (1998) continue the research initiated by Resnik. This work defines, for each verb of interest, a separate statistical model. Both the structure and the parameters of the models are inferred from the training data. The entire approach is grounded in fundamental information-theoretic principles.

For Li and Abe, a selectional preference model is a combination of a *cut* across the semantic class hierarchy and a probability distribution over the elements of the cut set. A cut establishes a partition of the set of WordNet's word senses (see Fig. 3). That is, a cut is a set of semantic classes that together cover all of the word senses such that each word sense belongs to exactly one of the classes of the set.[4] Associated with each concept in a cut is a probability. For example, if *food* is a member of the cut set, assigning it a probability of .6 is interpreted as indicating that 60% of the direct objects of the verb are expected to be food words.

The process for selecting the cut to be used for the model strives to balance two competing criteria: (i) that the model do a good job of predicting the actual data observed, and (ii) that the model be simple (with a small cut set). This balance is achieved by adhering to the minimum description length (MDL) principle (Rissanen, 1978). The MDL principle says that given a set of empirical observations, and a family of models under consideration, in choosing a model from the family, we should choose that model which enables us to describe the data most concisely. In information-theoretic terms, we are to choose that model which allows us to transmit, across a communication channel, information sufficient to reproduce the data at the other end, most concisely. The receptor, in order to reproduce the data, must be informed of the model chosen, and then, with knowledge of the model chosen, receive a description of the data.

Returning to our example in Fig. 3, if the cut for the verb *eat* were to include the FOOD concept, then the model would predict that all words under FOOD (e.g., *meat* and *beer*) are equally likely. If this is not too different from what is actually observed, then the cost of describing a more complex model, will not be offset by the gain in describing the data. Presumably this is not the case, and the data will show that word senses classified as ANIMAL-FLESH occur far more frequently than BEVERAGE word senses. There will be an increase in the length of the description of the model due to the increased number of parameters: there is one probability to be encoded for each concept in the cut. However, this increase will be more than offset by the decrease in the description of the data that results because of the improved fit of the model to what was actually observed.

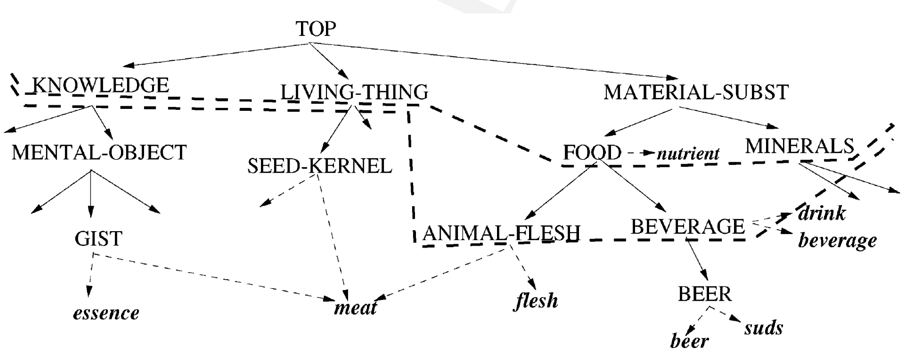

Fig. 3. Two example cuts for *eat*.

MDL is not the only way to decide on a cut. Clark and Weir (1999) describe a method where all the leaf classes of the hierarchy start in the cut and then the cut is moved up in the hierarchy, reasoning that lower-level sibling concepts (e.g., ANIMAL-FLESH and BEVERAGE) should be coalesced (into FOOD) if the probability of the occurrence of a FOOD word sense as direct object of the verb is independent of the subclass it belongs to. In this case, all the low-level probabilities are equal to each other and, hence, equal to the probability of seeing a word of the parent class. This decision is framed as hypothesis testing: the null hypothesis is that the probability of an element of the parent class is independent of whether it is an element of a particular child class. A χ^2 test is performed. If the result is significant, it is concluded that independence does not hold and the low-level semantic classes are used. Otherwise they are coalesced and the top-level is used.

One possible disadvantage of this approach, compared with MDL, is the arbitrary selection of the significance level used for the χ^2 test (.05 is used by Clark and Weir, 1999). On the other hand, this could be seen as an advantage, since it introduces a parameter that can be tuned for optimal performance for disparate tasks, different languages or different linguistic domains.

2.3. Hidden Markov models, Bayesian networks, and ambiguity in the training data

In this section, we present two further statistical models proposed for representing and inducing selectional preferences. In addition, we focus on handling word sense ambiguity in the training data.

The first model we present is that of Abney and Light (1999). In their approach each selectional preference (e.g., direct objects of *eat*) is represented as a separate HMM but all the HMMs have the same shape: the states and transitions of the HMMs are identified with the nodes and arcs of the given semantic class hierarchy (Fig. 4). The work described in the previous sections provides distributions over classes but is unclear as to how the models generate the words of the training data. It is simply assumed that all the words in a class are equally likely. In contrast, the HMMs allow different words of a class to have different probability distributions. Another attraction of the HMMs is that a number of interesting and useful distributions can be easily generated from them: the selectional preferences of a verb for its object can either

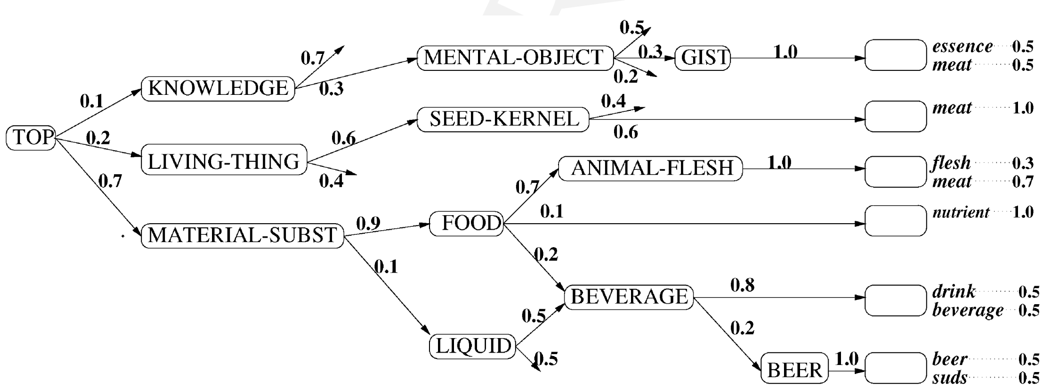

Fig. 4. An example HMM.

be a distribution over words, a distribution over word senses, or a distribution over semantic classes, all using the same underlying model.

Roughly speaking, an HMM is a stochastic version of a nondeterministic finite state machine. States change according to a state-specific distribution over the possible next states. A "run" of the type of HMM used by Abney and Light begins at the root of the semantic hierarchy. A transition from the current semantic class to a child class is chosen in accordance with the HMMs transition probabilities. This is done repeatedly until a terminal node (word sense) c is reached, at which point a word w is emitted in accordance with the probability of expressing sense c as word w. Hence, each HMM "run" can be identified with a path of arrows through the hierarchy of Fig. 4 from the root to a word sense, plus the word that was generated from the word sense; e.g., start at TOP, proceed to GIST and generate *essence*. Every observation sequence generated by the HMMs consists of a single noun: each run leads to a final state, at which point exactly one word is emitted. For these models there is not a word emission from each state visited. This is somewhat unusual, but formally speaking they are still HMMs, and the usual properties and algorithms apply.

Although the HMMs proposed by Abney and Light have many attractive features, successful parameter estimation proved elusive. In order to enable the parameter estimation algorithm (the forward–backward algorithm) to make generalizations rather than overfit the observed data, a bias towards a uniform distribution over state transitions is used. This bias is implemented by mixing in a uniform distribution when there is little evidence for a particular distribution.[5] This bias interacts with the topology of the semantic class hierarchy in problematic ways. Abney and Light describe a number of modifications to the parameter estimation algorithm that were helpful but ultimately unsuccessful. Thus, the potential of their approach has not yet been fully realized.

Ciaramita and Johnson (2000) follow Li and Abe in supposing that each verb selects for some set of WordNet classes as objects, and that the observed objects are indirect and noisy evidence of the selected classes. However, they ask not how strongly *eat* selects for FOOD (e.g., how often its direct objects are foods), but rather how likely it is that *eat* selects for FOOD at all. They treat this problem with Bayesian belief networks, allowing for an explicit and principled encoding of prior knowledge. The framework allows us to infer, for each class in the network, the probability that "the verb of interest, v, selects for the class, c." As usual, inference follows from a combination of the observed data and the knowledge encoded in the network.

The topology of Ciaramita and Johnson's Bayesian network is identical to that of WordNet. The probability distributions in the network are specified in accord with the following intuitions: (i) it is *a priori* unlikely that any given class will be selected for; (ii) a class is unlikely to be selected for if none of its parent classes is, but is likely to be selected for if at least one of its parent classes is; (iii) a word type is unlikely to appear in the corpus as direct object of the verb if none of its possible senses is selected for, but it is likely to appear at least once if at least one of its senses is selected for.

For a given verb, if it were known which of the top-level variables were *true*, i.e., which of the top-level classes were selected for, direct computation based on the "causal" knowledge encoded in the network could be performed to infer the probabilities that the verb selects for particular lower-level classes and appears with particular direct objects. Bayesian networks are designed, however, to allow inference in the other direction as well. In this scenario, it is the data that is observed, and probabilities for the possible values of (the latent) variables higher up in the network are inferred. When a word is observed to occur, it becomes more likely that one of its senses was selected for. A higher probability of preference for a class implies, in turn, higher probabilities of preference for classes which "cause" it to be preferred, as well as class nodes which it "causes" to be preferred.

As mentioned earlier, in the training data used here, an induction algorithm is not privy to the proper sense for the occurrences of ambiguous word types. In the work of both Resnik and Li and Abe, the counts for ambiguous words are spread evenly across the potential word senses. The hope is that, in general, the signal will be sufficiently strong to overcome the noise introduced by this approach. One would hope, for example, that there will be enough FOOD words appearing as the object of *eat* to overcome the effect of counting only some of the occurrences of *meat* as a FOOD, since *meat* can also be a MENTAL-OBJECT. Given the good results obtained by Resnik and Li and Abe, the signal does seem to be sufficiently strong. However, better performance may be possible if the problem of word sense ambiguity can be solved instead of ignored.

EM algorithms perform an iterative re-estimation of the parameters of a model in the face of "hidden" data (such as the word sense of a token). Both Abney and Light and Clark and Weir employed EM algorithms to their respective models. In addition, McCarthy (1997) applied re-estimation to the approach of Li and Abe. Intuitively, an EM algorithm starts with a guess at the proper model and uses this guess plus the training data to estimate the counts of the hidden word senses. These counts are then used to calculate the next model. The process is continued until the model no longer changes significantly.

Bayesian networks offer an alternative to dealing with the problem of incomplete data by exploiting a phenomenon which Pearl has called "explaining away" (Pearl, 1988). If an event is observed to occur (the alarm sounded), the probability for events that are possible causes (there was a burglar, the neighborhood cat was about) are increased. However, as evidence for one of the causes mounts, pressure for increasing the probability of alternate explanations is reduced. If "meow" is heard, the probability that the cat tripped the alarm increases. This decreases the probability that there was a burglar; the motivation for an increased probability of burglary having been *explained away*. If *meat* occurs as the object of *eat*, the probability that *eat* selects for ANIMAL-FLESH, SEED-KERNEL and GIST is raised for each. However, if many occurrences of other ANIMAL-FLESH and SEED-KERNEL words are observed, the probability that *eat* selects for these classes will be raised even further. This will be accompanied by a lower probability for the GISTconcept, and this lower probability will be accompanied by concomitant lower probabilities for its hypernyms. In this way an observation in one corner of the network ripples through the rest of the network.

3. Evaluation

In computational linguistics, formal evaluations provide a validation for a theory or approach. For many tasks, there exists a "gold standard" set of examples for which the outcome or answer has been generated by a human annotator. In many cases, multiple human annotators are used and the task is refined until inter-annotator agreement is acceptable (e.g., above 90%).

Table 3
Word sense disambiguation results

Method	
Random	28.5%
HMM (Abney & Light)	42.3%
Resnik	44.3%
Bayesian Belief network (Ciaramita & Johnson)	51.4%

For example, to evaluate part-of-speech tagging systems, one might give annotators a set of guidelines for hand-tagging a few thousand words of running text, and evaluate automatic systems on how well their tags matched the human ones. The inter-annotator agreement would serve as an upper bound on performance. An evaluation for selectional preferences along these lines would have humans generate selectional preferences for the test verbs and then score systems by how well they generated the same preferences. None of the work discussed here presents such an evaluation.

Another way of evaluating an induced set of selectional preferences is by showing their contribution to the performance of a related task. For example, word sense disambiguation results are reported by Resnik (1997), Abney and Light (1999), and Ciaramita and Johnson (2000). The training and test materials were extracted from the Penn Treebank syntactic parses of the Brown Corpus and the Semcor word sense data set. Semcor (which is distributed with WordNet) consists of 200,000 words of the Brown Corpus hand tagged with WordNet senses. Training data sets were then extracted for 100 verbs from the 800,000 words of the Brown Corpus that were *not* part of Semcor, using the Penn Treebank parses to find the heads of direct object complements. The test corpora were similarly extracted except that the Semcor portion of the Brown Corpus was used and the correct word sense of the object was noted. Each system was trained on the training set and then used to assign a word sense to the objects in the test set.

Table 3 presents the accuracy of each system on word sense disambiguation. The random method is simply to randomly pick a sense and is included as a baseline for comparison.

Other related task evaluations have also been performed. For example, Li and Abe evaluate their system on the task of prepositional phrase attachment.

In addition to direct evaluations and related task evaluations, selectional preferences can be evaluated as to how well they predict linguistic and psycholinguistic phenomena. Resnik (1996) shows that selectional association strength is predictive of implicit object alternations. In addition, he performed experiments comparing human plausibility judgments and his model's selectional preferences. The plausibility of direct objects such as *driver* and *engine* are compared in sentences such as *the mechanic warned the*... and a correlation between human and model plausibility ratings is shown to exist.

4. Conclusion

Resnik's dissertation (Resnik, 1993) initiated a new approach to selectional preference representation and induction. The approach combines knowledge represented in a pre-defined semantic class hierarchy with statistical tools including information theory, statistical modeling, and Bayesian inference. The final ingredient is a large corpus of written language from which to derive training materials. We have surveyed research that extends Resnik's initial work and discussed the strengths and weaknesses of each approach.

All of the approaches use a concept taxonomy to allow for generalizations that go beyond what could be inferred from the data alone. Dependence on a specified hierarchy also ensures that the selectional preference knowledge induced will be consistent with a given pre-conceived notion of what the semantic classes are. Further, all of the researchers have based their work on the WordNet semantic hierarchy—most surely because its coverage is extensive and it is readily available. There is nothing in these approaches, however, that is specific to WordNet, and all of them could work with other concept networks of a similar nature.

Notes

1. The work of Pustejovsky (1995) is a notable exception.
2. The predicate itself might have multiple senses and the different senses may have different preferences. For example, the verb *toast* would prefer newlyweds or breads depending on the sense being used. Again the work here does not take this issue into account. However, see (Agirre & Martinez, 2001) for work in this area.
3. This is so in Fig. 1 even though not all nine classes containing *meat* are mutually exclusive (*meat* is only three ways ambiguous).
4. For the purposes of their research, they treated the WordNet hierarchy as if it were a tree, although this is not quite accurate, since some WordNet classes do have multiple parents.
5. This method can also be seen as a Dirichlet prior. Being able to consider it as a prior results in the retention of the convergence characteristics of the relevant EM algorithm (Jason Eisner, personal communication).

Acknowledgments

The authors wish to express their gratitude to Steven Abney, Jason Eisner, and two anonymous reviewers for their valuable contributions to this paper.

References

Abney, S. (1997). Partial parsing via finite-state cascades. *Natural Language Engineering, 2* (4).

Abney, S., & Light, M. (1999). Hiding a semantic hierarchy in a Markov model. In *Proceedings of the ACL Workshop on Unsupervised Learning in Natural Language Processing*.

Agirre, E., & Martinez, D. (2001). Learning class-to-class selectional preferences. In *Proceedings of the ACL/EACL Workshop on Computational Natural Language Learning*.

Burnard, L. (1995). *Users reference guide for the British National Corpus*. Oxford University Computing Services.

Chomsky, N. (1965). *Aspects of the theory of syntax*. Cambridge, MA: MIT Press.

Ciaramita, M., & Johnson, M. (2000). Explaining away ambiguity: Learning verb selectional preference with Bayesian networks. In *Proceedings of the 18th International Conference on Computational Linguistics* (*COLING 2000*).

Clark, S., & Weir, D. (1999). An iterative approach to estimating frequencies over a semantic hierarchy. In P. Fung & J. Zhou (Eds.), *Proceedings of the 1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora* (pp. 258–265).

Fellbaum, C. (Ed.) (1998). *WordNet: An electronic lexical database*. Cambridge, MA: MIT Press.

Grishman, R., Hirschman, L., & Nhan, N. T. (1986). Discovery procedures for sublanguage selectional patterns: Initial experiments. *Computational Linguistics*, *12*(3), 205–215.

Katz, J. J., & Fodor, J. A. (1964). The structure of a semantic theory. In J. A. Fodor & J. J. Katz (Eds.), *The structure of language* (pp. 479–518). Englewood Cliffs, NJ: Prentice-Hall.

Lee, L., Pereira, F., & Tishby, N. (1993). Distributional clustering of English words. In *Proceedings of the 31th Annual Meeting of the Association for Computational Linguistics*.

Li, H., & Abe, N. (1998). Generalizing case frames using a thesaurus and the MDL principle. *Computational Linguistics*, *24*(2), 217–244.

Marcus, M., Santorini, B., & Marcinkiewicz, M. A. (1993). Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, *19*(2), 313–330.

McCarthy, D. (1997). Word sense disambiguation for acquisition of selectional preferences. In *Proceedings of the ACL/EACL Workshop on Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*.

Miller, G. (1990). WordNet: An on-line lexical database. *International Journal of Lexicography*, *3*(4).

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. San Mateo, CA: Morgan Kaufmann.

Pustejovsky, J. (1995). *The generative lexicon*. Cambridge, MA: MIT Press.

Resnik, P. (1993). *Selection and information: A class-based approach to lexical relationships*. Unpublished doctoral dissertation, University of Pennsylvania, Philadelphia, PA (available as Report 93-42 from the Institute for Research in Cognitive Science).

Resnik, P. (1996). Selectional constraints: An information-theoretic model and its computational realization. *Cognition*, *61*, 127–159.

Resnik, P. (1997). Selectional preference and sense disambiguation. In *Proceedings of the ANLP-97 workshop: Tagging text with lexical semantics: Why, what, and how*? Washington, DC.

Rissanen, J. (1978). Modeling by shortest data description. *Automatic*, *14*, 37–38.