A Class-based Probabilistic approach to Structural Disambiguation

Stephen Clark and David Weir
School of Cognitive and Computing Sciences
University of Sussex
Brighton, BN1 9HQ, UK
{stephecl,davidw}@cogs.susx.ac.uk

Abstract

Knowledge of which words are able to fill particular argument slots of a predicate can be used for structural disambiguation. This paper describes a proposal for acquiring such knowledge, and in line with much of the recent work in this area, a probabilistic approach is taken. We develop a novel way of using a semantic hierarchy to estimate the probabilities, and demonstrate the general approach using a prepositional phrase attachment experiment.

1 Introduction

Knowledge of which words are able to fill particular argument slots of a predicate can be used for structural disambiguation. In the following example (Charniak, 1993), the fact that *dog*, rather than *prize*, is often the subject of *run*, can be used to decide on the attachment site of the relative clause:

Fred awarded a prize for the dog that ran the fastest

We describe a proposal for acquiring such knowledge, and as in other recent work in this area (Resnik, 1993; Li and Abe, 1998), a probabilistic approach is taken. Using probabilities accords with the intuition that there are no absolute constraints on the arguments of predicates, but rather that constraints are satisfied to a certain degree (Resnik, 1993). Unfortunately, defining probabilities in terms of words leads to a model with a vast number of parameters, resulting in a sparse data problem. To overcome this, we propose to define a probability model in terms of senses from a semantic hierarchy, exploiting the fact that senses of nouns can be grouped together into semantically similar classes.

We use the semantic hierarchy of noun senses in WordNet (Fellbaum, 1998), which consists of 'lexicalised concepts' related by the 'is-a-kind-of' relation. If c' is a kind of c, then c is a *hypernym* of c', and c' a *hyponym* of c. Counts are passed up the hierarchy from the senses of nouns appearing in the data. Thus if *eat chicken* appears in the data, the count for this item passes up to $\langle$`meat`$\rangle$, $\langle$`food`$\rangle$, and all the other hypernyms of that sense of *chicken*.[1] In order to estimate the probability that a sense of *chicken* appears as the object of the verb *eat*, we represent $\langle$`chicken`$\rangle$ using a suitable hypernym, such as $\langle$`food`$\rangle$, and base our probability estimate on that instead. The level at which $\langle$`chicken`$\rangle$ is represented is crucial: it should be high enough for adequate counts to have accumulated, but not too high so that the hypernym is no longer representative of $\langle$`chicken`$\rangle$. An example of a hypernym which would be too high is $\langle$`entity`$\rangle$, as not all entities are semantically similar with respect to the object position of *eat*.

The problem of choosing an appropriate level in the hierarchy at which to represent a particular noun sense (given a predicate and argument position) has been investigated by Resnik (1993), Li and Abe (1998) and Ribas (1995). The learning mechanism presented here is a novel approach based on finding semantically similar sets of concepts in a hierarchy. We demonstrate the effectiveness of our approach using a PP-attachment experiment.

2 The Input Data and Semantic Hierarchy

The data used to estimate the probabilities is a multiset of 'co-occurrence triples': a noun

[1] We use italics when referring to words, and angled brackets for concepts. This notation does not always pick out a concept uniquely, but the context should make clear the concept being referred to.

lemma, verb lemma, and argument position.[2] Let the universe of verbs, argument positions and nouns that can appear in the input data be denoted $\mathcal{V} = \{ v_1, \ldots, v_{k_\mathcal{V}} \}$, $\mathcal{R} = \{ r_1, \ldots, r_{k_\mathcal{R}} \}$ and $\mathcal{N} = \{ n_1, \ldots, n_{k_\mathcal{N}} \}$, respectively. Such data can be obtained from a treebank, or from a shallow parser. Note that we do not distinguish between alternative senses of verbs, and assume that each instance of a noun in the data refers to exactly one concept.

The semantic hierarchy used is the noun hypernym taxonomy of WordNet (version 1.6).[3] Let $\mathcal{C} = \{ c_1, \ldots, c_{k_\mathcal{C}} \}$ be the set of concepts in WordNet ($k_\mathcal{C} \approx 66,000$). A concept is represented in WordNet by a synset: a set of synonymous words which can be used to denote that concept. For example, the concept 'cocaine', as in the drug, is represented by the following synset: $\{$*cocaine*, *cocain*, *coke*, *snow*, *C*$\}$. Let $\mathrm{syn}(c) \subseteq \mathcal{N}$ be the synset for the concept c, and let $\mathrm{cn}(n) = \{ c \mid n \in \mathrm{syn}(c) \}$ be the set of concepts that can be denoted by the noun n.

The hierarchy has the structure of a directed acyclic graph, although the number of nodes in the graph with more than one parent is only around one percent of the total. The edges in the graph form what we call the direct-isa relation (direct-isa $\subseteq \mathcal{C} \times \mathcal{C}$). Let isa = direct-isa* be the transitive, reflexive closure of direct-isa, so that $(c', c) \in \mathrm{isa} \Rightarrow c$ is a hypernym of c'; and let $\overline{c} = \{ c' \mid (c', c) \in \mathrm{isa} \}$ be the set consisting of the concept c and all of its hyponyms. Thus, the set $\overline{\langle\texttt{food}\rangle}$ contains all the concepts which are kinds of food, including $\langle\texttt{food}\rangle$.

Note that words in the data can appear in synsets anywhere in the hierarchy. Even concepts such as $\langle\texttt{entity}\rangle$, which appear near the root of the hierarchy, have synsets containing words which may appear in the data. The synset for $\langle\texttt{entity}\rangle$ is $\{$*entity*, *something*$\}$, and the words *entity* and *something* can appear in the argument positions of verbs in the data.

3 Probability Estimation

The problem being addressed in this section is to estimate $p(c|v,r)$, for $c \in \mathcal{C}$, $v \in \mathcal{V}$, and $r \in \mathcal{R}$. The probability $p(c|v,r)$ is the probability that some noun in $\mathrm{syn}(c)$, when denoting concept c, appears in position r of verb v (given r and v). Using the relative clause example from the introduction, the probabilities $p(\langle\texttt{dog}\rangle|run,\mathrm{subj})$ and $p(\langle\texttt{prize}\rangle|run,\mathrm{subj})$ can be compared to decide on the attachment site in *Fred awarded a prize for the dog that ran the fastest.* We expect $p(\langle\texttt{dog}\rangle|run,\mathrm{subj})$ to be greater than $p(\langle\texttt{prize}\rangle|run,\mathrm{subj})$. Although the focus is on $p(c|v,r)$, the techniques described here can be used to estimate other probabilities, such as $p(c,r|v)$. (In fact, the latter probability is used in the PP-attachment experiments described in Section 5.)

Using maximum likelihood to estimate $p(c|v,r)$ is not viable because of the huge number of parameters involved. Many combinations of c, v and r will not occur in the data. To reduce the number of parameters which need to be estimated, we utilise the fact that concepts can be grouped into classes, and represent c using a class $\overline{c'}$, for some hypernym c' of c. However, $p(\overline{c'}|v,r)$ cannot be used as an estimate of $p(c|v,r)$, as $p(\overline{c'}|v,r)$ is given by the following:

$$p(\overline{c'}|v,r) = \sum_{c'' \in \overline{c'}} p(c''|v,r)$$

The probability $p(\overline{c'}|v,r)$ increases as c' moves up the hierarchy. For example, $p(\overline{\langle\texttt{food}\rangle}|eat,\mathrm{obj})$ is not a good estimate of $p(\langle\texttt{chicken}\rangle|eat,\mathrm{obj})$. What can be done though, is to *condition* on sets of concepts, and use the probability $p(v|\overline{c'},r)$. If it can be shown that $p(v|\overline{c'},r)$, for some hypernym c' of c, is a reasonable estimate of $p(v|c,r)$, then we have a way of estimating $p(c|v,r)$. To get $p(v|c,r)$ from $p(c|v,r)$ Bayes rule is used:

$$p(c|v,r) = p(v|c,r)\frac{p(c|r)}{p(v|r)}$$

The probabilities $p(c|r)$ and $p(v|r)$ can be estimated using maximum likelihood estimates, as the conditioning event is likely to occur often enough for sparse data not to be a problem. (Alternatively one could back-off to $p(c)$ and $p(v)$ respectively, or use a linear combination of $p(c|r)$ and $p(c)$, and $p(v|r)$ and $p(v)$, respectively.) The formulae for these estimates will be given shortly. This only leaves $p(v|c,r)$. The

[2] Only verbs are considered here, but this work applies to other predicates which take arguments that can be organised into a semantic hierarchy.

[3] When we refer to concepts in WordNet, we mean concepts in WordNet's noun taxonomy.

proposal is to estimate $p(eat|\langle\texttt{chicken}\rangle, \text{obj})$ using $p(eat|\overline{\langle\texttt{food}\rangle}, \text{obj})$, or something similar. The following proposition shows that if $p(v|c'', r)$ is the same for each c'' in $\overline{c'}$, where c' is some hypernym of c, then $p(v|\overline{c'}, r)$ will be equal to $p(v|c, r)$:

$$p(v|c'', r) = k \text{ for all } c'' \in \overline{c'} \;\Rightarrow\; p(v|\overline{c'}, r) = k$$

The proof is as follows:

$$\begin{aligned} p(v|\overline{c'}, r) &= p(\overline{c'}|v, r)\frac{p(v|r)}{p(\overline{c'}|r)} \\ &= \frac{p(v|r)}{p(\overline{c'}|r)} \sum_{c'' \in \overline{c'}} p(c''|v, r) \\ &= \frac{p(v|r)}{p(\overline{c'}|r)} \sum_{c'' \in \overline{c'}} p(v|c'', r)\frac{p(c''|r)}{p(v|r)} \\ &= \frac{1}{p(\overline{c'}|r)}\, k \sum_{c'' \in \overline{c'}} p(c''|r) \\ &= k \end{aligned}$$

So in order to estimate $p(v|c, r)$, we need a way of searching for a set $\overline{c'}$, where c' is a hypernym of c, which consists of concepts c'' which have similar $p(v|c'', r)$. Of course we cannot expect to find a set consisting of concepts which have identical $p(v|c'', r)$, which the proposition strictly requires, but if the $p(v|c'', r)$ are similar, then we can expect $p(v|\overline{c'}, r)$ to be a reasonable estimate of $p(v|c, r)$. We refer to the set $\overline{c'}$ as the 'similarity-class' of c, and the suitable hypernym, c', as $\text{top}(c, v, r)$. The next section explains how we determine similarity classes. The maximum likelihood estimates for the relevant probabilities are given in Table 1.[4]

4 Finding Similarity-classes

First we explain how we determine if a set of concepts has similar $p(v|c'', r)$ for each concept c'' in the set. Then we explain how we determine $\text{top}(c, v, r)$.

[4] Since we are assuming the data is not sense disambiguated, $\text{freq}(c, v, r)$ cannot be obtained by simply counting senses. The standard approach, which is adopted here, is to estimate $\text{freq}(c, v, r)$ by distributing the count for each noun n in $\text{syn}(c)$ evenly among all senses of the noun. Yarowsky (1992) and Resnik (1993) explain how the noise introduced by this technique tends to dissipate as counts are passed up the hierarchy.

Table 1: Maximum Likelihood Estimates – $\text{freq}(c, v, r)$ is the number of (n, v, r) triples in the data in which n is being used to denote c.

$$\hat{p}(c|r) = \frac{\text{freq}(c,r)}{\text{freq}(r)} = \frac{\sum_{v' \in \mathcal{V}} \text{freq}(c,v',r)}{\sum_{v' \in \mathcal{V}} \sum_{c' \in \mathcal{C}} \text{freq}(c',v',r)}$$

$$\hat{p}(v|r) = \frac{\text{freq}(v,r)}{\text{freq}(r)} = \frac{\sum_{c' \in \mathcal{C}} \text{freq}(c',v,r)}{\sum_{v' \in \mathcal{V}} \sum_{c' \in \mathcal{C}} \text{freq}(c',v',r)}$$

$$\hat{p}(v|\overline{c'}, r) = \frac{\text{freq}(\overline{c'},v,r)}{\text{freq}(\overline{c'},r)} = \frac{\sum_{c'' \in \overline{c'}} \text{freq}(c'',v,r)}{\sum_{v' \in \mathcal{V}} \sum_{c'' \in \overline{c'}} \text{freq}(c'',v',r)}$$

The method used for comparing the $p(v|c'', r)$ for c'' in some set $\overline{c'}$, is based on the technique in Clark and Weir (1999) used for finding homogeneous sets of concepts in the WordNet noun hierarchy. Rather than directly compare estimates of $p(v|c'', r)$, which are likely to be unreliable, we consider the children of c', and use estimates based on counts which have accumulated at the children. If c' has children $c'_1, c'_2, \ldots, c'_n$, we compare $p(v|\overline{c'_i}, r)$ for each i. This is an approximation, but if the $p(v|\overline{c'_i}, r)$ are similar, then we assume that the $p(v|c'', r)$ for c'' in $\overline{c'}$ are similar too.

To determine whether the children of some hypernym c' have similar $p(v|\overline{c'_i})$, where c'_i is the ith child, we apply a χ^2 test to a contingency table of frequency counts. Table 2 shows some example frequencies for c' equal to $\langle\texttt{nutriment}\rangle$, in the object position of *eat*. The figures in brackets are the expected values, based on the marginal totals in the table. The null hypothesis of the test is that $p(v|\overline{c'_i}, r)$ is the same for each i. For Table 2 the null hypothesis is that for every child, c'_i, of $\langle\texttt{nutriment}\rangle$, the probability $p(eat|\overline{c'_i}, \text{obj})$ is the same.

The log-likelihood χ^2 statistic corresponding to Table 2 is 4.8. The log-likelihood χ^2 statistic is used rather than the Pearson's χ^2 statistic because it is thought to be more appropriate when the counts in the contingency table are low (Dunning, 1993). This tends to occur when the test is being applied to a set of concepts near the foot of the hierarchy.[5] We compared

[5] Fisher's exact test could be used for tables with low counts, but we do not do so because tables dominated by low counts are likely to have a high percentage of noise, due to the way counts for a noun are split among

Table 2: Contingency table for children of ⟨`nutriment`⟩

$\overline{c_i}$	$\hat{\mathrm{freq}}(\overline{c_i}, eat, \mathrm{obj})$		$\hat{\mathrm{freq}}(\overline{c_i}, \mathrm{obj}) - \hat{\mathrm{freq}}(\overline{c_i}, eat, \mathrm{obj})$		$\hat{\mathrm{freq}}(\overline{c_i}, \mathrm{obj}) = \sum_{v \in \mathcal{V}} \hat{\mathrm{freq}}(\overline{c_i}, v, \mathrm{obj})$
$\overline{\langle\texttt{milk}\rangle}$	0.0	(0.6)	9.0	(8.4)	9.0
$\overline{\langle\texttt{meal}\rangle}$	8.5	(5.6)	78.0	(80.9)	86.5
$\overline{\langle\texttt{course}\rangle}$	1.3	(1.7)	24.7	(24.3)	26.0
$\overline{\langle\texttt{dish}\rangle}$	5.3	(5.7)	82.3	(81.9)	87.6
$\overline{\langle\texttt{delicacy}\rangle}$	0.3	(1.8)	27.4	(25.9)	27.7
	15.4		221.4		236.8

the performance of log-likelihood χ^2 and Pearson's χ^2 using the PP-attachment experiment described in Section 5. It was found that the log-likelihood χ^2 test did perform slightly better. For a significance level of 0.05 (which is the level used in the experiments), with 4 degrees of freedom, the critical value is 14.86 (Howell, 1997). Thus in this case, the null hypothesis would not be rejected.

In order to determine $\mathrm{top}(c, v, r)$, we compare $p(v|\overline{c_i}, r)$ for the children of the hypernyms of c. Initially $\mathrm{top}(c, v, r)$ is assigned to be the concept c itself. Then, by working up the hierarchy, $\mathrm{top}(c, v, r)$ is reassigned to be successive hypernyms of c until the siblings of $\mathrm{top}(c, v, r)$ have significantly different probabilities. In cases where a concept has more than one parent, the parent is chosen which results in the lowest χ^2 value as this indicates the $p(v|\overline{c_i}, r)$ are more similar. The set $\overline{\mathrm{top}(c, v, r)}$ is the similarity-class of c for verb v and position r.

The next section provides evidence that the technique for choosing $\mathrm{top}(c, v, r)$, which we call the 'similarity-class' technique, does select an appropriate level of generalisation.

5 Experiments using PP-attachment ambiguity

The PP-attachment problem we address considers 4-tuples of the form v, n_1, pr, n_2, and the problem is to decide whether the prepositional phrase pr n_2 attaches to the verb v or the noun n_1. For example, in the following case the problem is to decide whether *from minister* attaches to *await* or *approval*:

await approval from minister

We chose the PP-attachment problem because PP-attachment is a pervasive form of ambiguity, and there exist standard training and test data, which makes for easy comparisons with other approaches. This problem has been tackled by a number of researchers. Brill and Resnik (1994), Ratnaparkhi et al. (1994), Collins (1995), Zavrel and Daelemans (1997) all report results between 81% and 85%, with Stetina and Nagao (1997) reporting a result of 88%, which matches the human performance on this task reported by Ratnaparkhi et al. (1994).

Although the PP-attachment problem has characteristics that make it suitable for evaluation, it presents a much bigger sparse data problem than would be expected in other problems such as relative clause attachment. The reason for this is that we need to consider how a concept is associated with *combinations* of predicates and prepositions. The approach described here uses probabilities of the form $p(c, pr|v)$ and $p(c, pr|n_1)$, where $c \in \mathrm{cn}(n_2)$. This means that for many predicate/preposition combinations which occur infrequently in the data, there are few examples of n_2 which can be used for populating WordNet in these cases. Despite this, we were still able to carry out an evaluation by considering subsets of the test data for which the relevant predicate/preposition combinations did occur frequently in the training data.

We decide on the attachment site by compar-

alternative senses. We rely on the log-likelihood χ^2 test returning a non-significant result in these cases.

ing $p(c_v, pr|v)$ and $p(c_{n_1}, pr|n_1)$, where

$$c_v = \arg\max_{c \in \mathrm{cn}(n_2)} p(c, pr|v)$$
$$c_{n_1} = \arg\max_{c \in \mathrm{cn}(n_2)} p(c, pr|n_1)$$

The sense of n_2 is chosen which maximises the relevant probability in each potential attachment case. If $p(c_v, pr|v)$ is greater than $p(c_{n_1}, pr|n_1)$, the attachment is made to v, otherwise to n_1. If n_2 is not in WordNet we compare $p(pr|v)$ and $p(pr|n_1)$. Probabilities of the form $p(c, pr|v)$ and $p(c, pr|n_1)$ are used rather than $p(c|v, pr)$ and $p(c|n_1, pr)$, because the association between the preposition and v and n_1 contains useful information. In fact, for a lot of cases this information alone can be used to decide on the correct attachment site. The original corpus-based method of Hindle and Rooth (1993) used exactly this information. Thus the method described here can be thought of as Hindle and Rooth's method with additional class-based information about n_2.

In order to estimate $p(c_v, pr|v)$ (and $p(c_{n_1}, pr|n_1)$) we apply the same procedure as described in Section 3, first rewriting the probability using Bayes' rule:

$$\begin{aligned} p(c_v, pr|v) &= p(v|c_v, pr)\frac{p(c_v, pr)}{p(v)} \\ &= p(v|c_v, pr)\frac{p(pr|c_v)p(c_v)}{p(v)} \end{aligned}$$

The probabilities $p(c_v)$ and $p(v)$ can be estimated using maximum likelihood estimates, and $p(v|c_v, pr)$ and $p(pr|c_v)$ can be estimated using maximum likelihood estimates of $p(v|\overline{\mathrm{top}(c_v, v, pr)}, pr)$ and $p(pr|\overline{\mathrm{top}(c_v, pr)})$ respectively.[6]

We used the training and test data described in Ratnaparkhi et al. (1994), which was taken from the Penn Treebank and has now become the standard data set for this task. The data set consists of tuples of the form (v, n_1, pr, n_2), together with the attachment site for each tuple. There is also a development set to prevent implicit training on the test set during development. We extracted (v, pr, n_2) and (n_1, pr, n_2) triples from the training set, and in order to increase the number of training triples, we also extracted triples from unambiguous cases of attachment in the Penn Treebank. We preprocessed the training and test data by lemmatising the words, replacing numerical amounts with the words 'definite_quantity', replacing monetary amounts with the words 'sum_of_money' etc. We then ignored those triples in the resulting training set (but not test set) for which n_2 was not in WordNet, which left a total of $66,881$ triples of training data. The test set contains $3,097$ examples.

Table 3 gives some examples of the extent to which the similarity-class technique is generalising, using the training data just described, and a significance level of 0.05. The chosen hypernym is shown in upper case. Note that the WordNet hierarchy consists of nine separate sub-hierarchies, headed by such concepts as ⟨`entity`⟩, ⟨`abstraction`⟩, ⟨`psychological_feature`⟩, but we assume the existence of a single root which dominates each of the sub-hierarchies, which is referred to as ⟨`root`⟩. In cases where WordNet is very sparsely populated, it is preferable to go to ⟨`root`⟩, rather than stay at the root of one of the sub-hierarchies where the data may be noisy or too sparse to be of any use. The table shows that with the amount of data available from the Treebank, the similarity-class technique is selecting a level at or close to ⟨`root`⟩ in many cases.

We compared the similarity-class technique with fixing the level of generalisation. Two fixed levels were used: the root of the entire hierarchy (⟨`root`⟩), and the set consisting of the roots of each of the 9 sub-hierarchies. The procedure which always selects ⟨`root`⟩ ignores any information about n_2, and is equivalent to comparing $p(pr|v)$ and $p(pr|n_1)$, which is the Hindle and Rooth approach. The results on the $3,097$ test cases are shown in Table 4. We used a significance level α of 0.05 for the χ^2 test.[7]

As the table shows, the disambiguation accuracy is below the state of the art. However, the results are comparable with those of Li and

[6] In Section 4 we only gave the procedure for determining $\mathrm{top}(c_v, v, pr)$, but $\mathrm{top}(c_v, pr)$ can be determined in an analogous fashion.

[7] Similar results were obtained using alternative levels of significance. Rather than simply selecting a value for α such as 0.05, α can be treated as a parameter of the model, whose optimum value can be obtained by running the disambiguation method on some held-out supervised data.

Table 3: How the similarity-class technique chooses top(c, v, pr) and top(c, n_1, pr)

(n_1, pr, c)	Hypernyms of c
(*bid,for,*⟨`company`⟩)	⟨`company`⟩⟨`establishment`⟩⟨`organisation`⟩⟨`social_group`⟩⟨`GROUP`⟩⟨`root`⟩
(*concern,about,*⟨`risk`⟩)	⟨`risk`⟩⟨`venture`⟩⟨`task`⟩⟨`work`⟩⟨`activity`⟩⟨`act`⟩⟨`ROOT`⟩
(*billion,in,*⟨`cash`⟩)	⟨`cash`⟩⟨`currency`⟩⟨`monetary_system`⟩⟨`asset`⟩⟨`POSSESSION`⟩⟨`root`⟩
(v, pr, c)	
(*notify,of,*⟨`transaction`⟩)	⟨`transaction`⟩⟨`group_action`⟩⟨`act`⟩⟨`ROOT`⟩
(*close,at,*⟨`definite_quantity`⟩)	⟨`DEFINITE_QUANTITY`⟩⟨`measure`⟩⟨`abstraction`⟩⟨`root`⟩
(*meet,with,*⟨`official`⟩)	⟨`official`⟩⟨`adjudicator`⟩⟨`person`⟩⟨`life_form`⟩⟨`CAUSAL_AGENT`⟩⟨`entity`⟩⟨`root`⟩

Table 4: Complete test set –3097 test cases

Generalisation technique	% correct
Similarity-class	80.3
Select root of sub-hierarchy	77.9
Always select ⟨`root`⟩	79.0

Abe (1998) who adopt a similar approach using WordNet, but with a different training and test set. Li and Abe improved on the Hindle and Rooth technique by 1.5%, which is in line with our results. As an evaluation of the similarity-class technique, the result is inconclusive. The reason for this is that when the technique was being used to estimate $p(v|c_v, pr)$ and $p(n_1|c_{n_1}, pr)$, in many cases the root of the hierarchy was being chosen as the appropriate level of generalisation, due to a sparsely populated WordNet in that instance. Recall that this is largely due to the fact that we are attempting to populate WordNet for combinations of predicates and prepositions. In such cases the similarity-class technique is not helping because there is very little or no information about n_2.[8]

[8] In an effort to obtain more data we applied the extraction heuristic of Ratnaparkhi (1998) to Wall Street Journal text, which increased the number of training triples by a factor of 10. This only achieved comparable results, however, presumably because the high volume of noise in the data outweighs the benefit of the increase in data size. Ratnaparkhi reports only 69% accuracy for the extraction heuristic when applied to the Penn Treebank (excluding cases where the preposition is *of*).

Table 5: ⟨`root`⟩ being selected for both attachment points – 113 test cases

Generalisation technique	% correct
Similarity-class	90.3
Select root of sub-hierarchy	81.4
Always select ⟨`root`⟩	79.6

Table 6: ⟨`root`⟩ being selected for at most one of the attachment points – 1032 test cases

Generalisation technique	% correct
Similarity-class	88.1
Select root of sub-hierarchy	85.5
Always select ⟨`root`⟩	85.5

In order to evaluate the similarity-class technique further, we took those test cases for which the root was not being selected when estimating both $p(v|c_v, pr)$ and $p(n_1|c_{n_1}, pr)$. This applied to 113 cases. The results are given in Table 5. We also took those test cases for which the root was being selected when estimating at most one of $p(v|c_v, pr)$ and $p(n_1|c_{n_1}, pr)$. This applied to 1032 test cases. The results are shown in Table 6.

6 Conclusions

We have shown that when instances of WordNet are well populated with examples of n_2, the method described here for solving PP-attachment ambiguities is highly accurate. When WordNet is sparsely populated, the method automatically resorts to comparing just the preposition and each of the potential attachment sites, as the similarity-class technique will select $\langle\texttt{root}\rangle$ as the appropriate level of generalisation for n_2 in such cases. We have also shown the similarity-class technique to be superior to using a fixed level of generalisation in WordNet.

Further work will look at how to integrate probabilities such as $p(c|v,r)$ into a model of dependency structure, similar to that of Collins (1996) and Collins (1997), which can be used for parse selection. However, knowledge of selectional preferences cannot by itself solve the problem of structural disambiguation, and this further work will also look at using additional knowledge, such as subcategorisation information.

References

Eric Brill and Philip Resnik. 1994. A rule-based approach to prepositional phrase attachment disambiguation. In *Proceedings of the fifteenth International Conference on Computational Linguistics*.

Eugene Charniak. 1993. *Statistical Language Learning*. The MIT Press.

Stephen Clark and David Weir. 1999. An iterative approach to estimating frequencies over a semantic hierarchy. In *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, pages 258–265.

Michael Collins. 1995. Prepositional phrase attachment through a backed-off model. In *Proceedings of the Third Workshop on Very Large Corpora*, pages 27–38, Cambridge, Massachusetts.

Michael Collins. 1996. A new statistical parser based on bigram lexical dependencies. In *Proceedings of the 34th Annual Meeting of the ACL*, pages 184–191.

Michael Collins. 1997. Three generative, lexicalised models for statistical parsing. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics*, pages 16–23.

Ted Dunning. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1):61–74.

Christiane Fellbaum, editor. 1998. *WordNet An Electronic Lexical Database*. The MIT Press.

Donald Hindle and Mats Rooth. 1993. Structural ambiguity and lexical relations. *Computational Linguistics*, 19(1):103–120.

David Howell. 1997. *Statistical Methods for Psychology: 4th ed.* Duxbury Press.

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

Adwait Ratnaparkhi, Jeff Reynar, and Salim Roukos. 1994. A maximum entropy model for prepositional phrase attachment. In *Proceedings of the ARPA Human Language Technology Workshop*, pages 250–255.

Adwait Ratnaparkhi. 1998. Unsupervised statistical models for prepositional phrase attachment. In *Proceedings of the Seventeenth International Conference on Computational Linguistics*, Montreal, Canada, Aug.

Philip Resnik. 1993. *Selection and Information: A Class-Based Approach to Lexical Relationships.* Ph.D. thesis, University of Pennsylvania.

Francesc Ribas. 1995. On learning more appropriate selectional restrictions. In *Proceedings of the Seventh Conference of the European Chapter of the Association for Computational Linguistics*, Dublin, Ireland.

Jiri Stetina and Makoto Nagao. 1997. Corpus based PP attachment ambiguity resolution with a semantic dictionary. In *Proceedings of the Fifth Workshop on Very Large Corpora*, pages 66–80, Beijing and Hong Kong.

David Yarowsky. 1992. Word-sense disambiguation using statistical models of Roget's categories trained on large corpora. In *Proceedings of COLING-92*, pages 454–460.

Jakub Zavrel and Walter Daelemans. 1997. Memory-based learning: Using similarity for smoothing. In *Proceedings of ACL/EACL-97*, Madrid, Spain.