Class-Based Probability Estimation using a Semantic Hierarchy

Stephen Clark[*]
University of Edinburgh

David Weir[†]
University of Sussex

This paper concerns the estimation of a particular kind of probability, namely the probability of a noun sense appearing as a particular argument of a predicate. In order to overcome the accompanying sparse data problem, the proposal is to define the probabilities in terms of senses from a semantic hierarchy, and exploit the fact that the senses can be grouped into classes consisting of semantically similar senses. There is a particular focus on the problem of how to determine a suitable class for a given sense, or, alternatively, how to determine a suitable level of generalisation in the hierarchy. A procedure is developed which uses a chi-squared test to determine a suitable level. In order to test the performance of the estimation method, a pseudo disambiguation task is used, together with two alternative estimation methods. Each method uses a different generalisation procedure; the first alternative uses the Minimum Description Length principle, and the second uses Resnik's measure of selectional preference. In addition, the performance of our method is investigated using both the standard Pearson chi-squared statistic and the log-likelihood chi-squared statistic.

1 Introduction

This paper[1] concerns the problem of how to estimate the probabilities of noun senses appearing as particular arguments of predicates. Such probabilities can be useful for a variety of NLP tasks, such as structural disambiguation and statistical parsing, word sense disambiguation, anaphora resolution and language modelling. To see how such knowledge can be used to resolve structural ambiguities, consider the following prepositional phrase attachment ambiguity:

Example 1
Fred ate strawberries with a spoon.

The ambiguity arises because the prepositional phrase *with a spoon* can attach to either *strawberries* or *ate*. The ambiguity can be resolved by noting that the correct sense of *spoon* is more likely to be an argument of "*ate-with*" than "*strawberries-with*" (Li and Abe, 1998; Clark and Weir, 2000).

The problem with estimating a probability model defined over a large vocabulary of predicates and noun senses is that this involves a huge number of parameters, which results in a sparse data problem. In order to reduce the number of parameters, we pro-

∗ Division of Informatics, University of Edinburgh, 2 Buccleuch Place, Edinburgh, EH8 9LW, UK. E-mail: stephenc@cogsci.ed.ac.uk
† School of Cognitive and Computing Sciences, University of Sussex, Brighton, BN1 9QH, UK. E-mail: david.weir@cogs.susx.ac.uk
1 This is an extended and updated version of a paper that appeared in the proceedings of NAACL 2001.

pose to define a probability model over senses in a semantic hierarchy, and exploit the fact that senses can be grouped into classes consisting of semantically similar senses. The assumption underlying this approach is that the probability of a noun sense can be approximated by a probability based on a suitably chosen class. For example, it seems reasonable to suppose that the probability of (the food sense of) *chicken* appearing as an object of the verb *eat* can be approximated in some way by a probability based on a class such as FOOD.

There are two elements to the problem of using a class to estimate the probability of a noun sense. First, given a suitably chosen class, how can that class be used to estimate the probability of the sense? And second, given a noun sense, how can a suitable class be determined? This work offers novel solutions to both problems, and there is a particular focus on the second question, which can be thought of as how to find a suitable level of generalisation in the hierarchy.[2]

The semantic hierarchy used is the noun hierarchy of WordNet (Fellbaum, 1998), version 1.6. Previous work has considered how to estimate probabilities using classes from WordNet, in the context of acquiring selectional preferences (Resnik, 1998; Ribas, 1995; Li and Abe, 1998; McCarthy, 2000), and this previous work has also addressed the question of how to determine a suitable level of generalisation. Li and Abe use the Minimum Description Length principle to obtain a level of generalisation, and Resnik uses a simple technique based on a statistical measure of selectional preference. (The work by Ribas builds on that by Resnik, and the work by McCarthy builds on that by Li and Abe.) We compare our estimation method with those of Resnik and Li and Abe, using a pseudo disambiguation task. Our method outperforms these alternatives on the pseudo disambiguation task, and an analysis of the results shows that the generalisation methods of Resnik and Li and Abe appear to be over-generalising, at least for this task.

Note that the problem being addressed here is the engineering problem of estimating predicate-argument probabilities, with the aim of producing estimates that will be useful for NLP applications. In particular, we are not addressing the problem of acquiring selectional restrictions in the way this is usually construed (Resnik, 1993; Ribas, 1995; McCarthy, 1997; Li and Abe, 1998; Wagner, 2000). The purpose of using a semantic hierarchy for generalisation is to overcome the sparse data problem, rather than find a level of abstraction that best represents the selectional restrictions of some predicate. This point is considered further in Section 5.

The next section describes the noun hierarchy from WordNet, and gives a more precise description of the probabilities to be estimated. Section 3 shows how a class from WordNet can be used to estimate the probability of a noun sense. Section 4 shows how a chi-squared test is used as part of the generalisation procedure, and Section 5 describes the generalisation procedure. Section 6 describes the alternative class-based estimation methods used in the pseudo disambiguation experiments, and Section 7 presents those experiments.

2 The Semantic Hierarchy

The noun hierarchy of WordNet consists of senses, or what Miller (1998) calls *lexicalised concepts*, organised by the "is-a-kind-of" relation. Note that we are using *concept* to refer to a lexicalised concept or sense, and not a set of senses; we use *class* to refer to a set of senses. There are around $66,000$ different concepts in the noun hierarchy of version

2 A third element of the problem, namely obtaining arguments of predicates as training data, is not considered here. We assume the existence of such data, obtained from a treebank or shallow parser.

1.6. A concept in WordNet is represented by a "synset", which is the set of synonymous words that can be used to denote that concept. For example, the synset for the concept $\langle\text{cocaine}\rangle$[3] is $\{\, cocaine, cocain, coke, snow, \mathsf{C} \,\}$. Let $\mathsf{syn}(c)$ be the synset for concept c, and let $\mathsf{cn}(n) = \{\, c \mid n \in \mathsf{syn}(c) \,\}$ be the set of concepts that can be denoted by noun n.

The hierarchy has the structure of a directed acyclic graph (although only around 1% of the nodes have more than one parent), where the edges of the graph constitute what we call the "$\mathsf{direct} - \mathsf{isa}$" relation. Let isa be the transitive, reflexive closure of $\mathsf{direct} - \mathsf{isa}$, then $c' \ \mathsf{isa}\ c$ implies c' is a kind of c. If $c' \ \mathsf{isa}\ c$, then c is a *hypernym* of c' and c' is a *hyponym* of c. In fact, the hierarchy is not a single hierarchy, but consists of nine separate sub-hierarchies, headed by the most general kind of concept, such as $\langle\text{entity}\rangle$, $\langle\text{abstraction}\rangle$, $\langle\text{event}\rangle$, $\langle\text{psychological\_feature}\rangle$. For the purposes of this work we add a common root dominating the nine sub-hierarchies, which we denote $\langle\text{root}\rangle$.

There are some important points of clarification regarding the hierarchy. First, every concept has a non-empty synset (except the notional concept $\langle\text{root}\rangle$). Even the most general concepts, such as $\langle\text{entity}\rangle$, can be denoted by some noun; the synset for $\langle\text{entity}\rangle$ is $\{\, entity, something \,\}$. Second, there is an important distinction between an individual concept and a set of concepts. For example, the individual concept $\langle\text{entity}\rangle$ should not be confused with the set or class consisting of concepts denoting kinds of entities. To make this distinction clear, we use $\overline{c} = \{\, c' \mid c' \ \mathsf{isa}\ c \,\}$ to denote the set of concepts dominated by concept c, including c itself. For example, $\overline{\langle\text{animal}\rangle}$ is the set consisting of those concepts corresponding to kinds of animals (including $\langle\text{animal}\rangle$ itself).

The probability of a concept appearing as an argument of a predicate is written $p(c|v, r)$, where c is a concept in WordNet, v is a predicate and r is an argument position.[4] The focus in this paper is on the arguments of verbs, but the techniques can be applied to any predicate that takes nominal arguments, such as adjectives. The probability $p(c|v, r)$ is to be interpreted as follows: this is the probability that some noun n in $\mathsf{syn}(c)$, when denoting concept c, appears in position r of verb v (given v and r). The example used throughout the paper is $p(\langle\text{dog}\rangle|run, \text{subj})$, which is the conditional probability that some noun in the synset of $\langle\text{dog}\rangle$, when denoting the concept $\langle\text{dog}\rangle$, appears in the subject position of the verb *run*. Note that, in practice, no distinction is made between the different senses of a verb (although the techniques do allow such a distinction), and each use of a noun is assumed to correspond to exactly one concept.[5]

3 Class-Based Probability Estimation

This section explains how a set of concepts, or class, from WordNet can be used to estimate the probability of an individual concept. More specifically, we explain how a set of concepts $\overline{c'}$, where c' is some hypernym of concept c, can be used to estimate $p(c|v, r)$. (Recall that $\overline{c'}$ denotes the set of concepts dominated by c', including c' itself.) One possible approach would be to simply substitute $\overline{c'}$ for the individual concept c. However, this is a poor solution, since $p(\overline{c'}|v, r)$ is the conditional probability that some noun denoting a concept in $\overline{c'}$ appears in position r of verb v. For example, $p(\overline{\langle\text{animal}\rangle}|run, \text{subj})$ is the probability that some noun denoting a kind of animal appears in the subject position of the verb *run*. Probabilities of sets of concepts are obtained by summing over the

3 Angled brackets are used to denote concepts in the hierarchy.
4 The term "predicate" is used loosely here, in that the predicate does not have to be a semantic object, but can simply be a word form.
5 A recent paper which extends the acquisition of selectional preferences to sense-sense relationships is Agirre and Martinez (2001).

concepts in the set:

$$p(\overline{c'}|v,r) = \sum_{c'' \in \overline{c'}} p(c''|v,r) \tag{1}$$

This means that $p(\overline{\langle \text{animal} \rangle}|run, \text{subj})$ is likely to be much greater than $p(\langle \text{dog} \rangle|run, \text{subj})$, and not a good approximation of $p(\langle \text{dog} \rangle|run, \text{subj})$.

What can be done, though, is to *condition* on sets of concepts. If it can be shown that $p(v|\overline{c'}, r)$, for some hypernym c' of c, is a reasonable approximation of $p(v|c,r)$, then we have a way of estimating $p(c|v,r)$. The probability $p(v|c,r)$ can be obtained from $p(c|v,r)$ using Bayes' theorem:

$$p(c|v,r) = p(v|c,r)\frac{p(c|r)}{p(v|r)} \tag{2}$$

Since $p(c|r)$ and $p(v|r)$ are conditioned on the argument slot only, we assume these can be estimated satisfactorily using relative frequency estimates. Alternatively, a standard smoothing technique such as Good-Turing could be used.[6] This leaves $p(v|c,r)$. Continuing with the $\langle \text{dog} \rangle$ example, the proposal is to estimate $p(run|\langle \text{dog} \rangle, \text{subj})$ using a relative frequency estimate of $p(run|\overline{\langle \text{animal} \rangle}, \text{subj})$, or an estimate based on a similar, suitably chosen class. Thus, assuming this choice of class, $p(\langle \text{dog} \rangle|run, \text{subj})$ would be approximated as follows:

$$p(\langle \text{dog} \rangle|run, \text{subj}) \approx p(run|\overline{\langle \text{animal} \rangle}, \text{subj})\frac{p(\langle \text{dog} \rangle|\text{subj})}{p(run|\text{subj})} \tag{3}$$

The following derivation shows that if $p(v|\overline{c'_i}, r) = k$ for each child c'_i of c', and $p(v|c', r) = k$, then $p(v|\overline{c'}, r)$ is also equal to k:

$$\begin{aligned}
p(v|\overline{c'}, r) &= p(\overline{c'}|v,r)\frac{p(v|r)}{p(\overline{c'}|r)} && (4)\\
&= \frac{p(v|r)}{p(\overline{c'}|r)}\left(\sum_i p(\overline{c'_i}|v,r) + p(c'|v,r)\right) && (5)\\
&= \frac{p(v|r)}{p(\overline{c'}|r)}\left(\sum_i p(v|\overline{c'_i}, r)\frac{p(\overline{c'_i}|r)}{p(v|r)} + p(v|c', r)\frac{p(c'|r)}{p(v|r)}\right) && (6)\\
&= \frac{1}{p(\overline{c'}|r)}\left(\sum_i k\, p(\overline{c'_i}|r) + k\, p(c'|r)\right) && (7)\\
&= \frac{k}{p(\overline{c'}|r)}\left(\sum_i p(\overline{c'_i}|r) + p(c'|r)\right) && (8)\\
&= k && (9)
\end{aligned}$$

Note that the proof only applies to a tree, since the proof assumes that $\overline{c'}$ is partitioned by c' and the sets of concepts dominated by each of the daughters of c', which is not necessarily true for a DAG. WordNet is a DAG, but is a close close approximation to a tree, and so we assume this will not be a problem in practice.[7]

6 Unsmoothed estimates were used in this work.

7 Li and Abe (1998) also develop a theoretical framework that only applies to a tree, and turn WordNet into a tree by copying each subgraph with multiple parents. One way to extend the experiments in Section 7 would be to investigate whether this transformation has an impact on the results.

The previous derivation shows how probabilities conditioned on sets of concepts can remain constant when moving up the hierarchy, and this suggests a way of finding a suitable set, $\overline{c'}$, as a generalisation for concept c: initially set c' equal to c, and move up the hierarchy, changing the value of c', until there is a significant change in $p(v|\overline{c'}, r)$. Estimates of $p(v|\overline{c'_i}, r)$, for each child c'_i of c', can be compared to see if $p(v|\overline{c'}, r)$ has significantly changed. (We ignore the probability $p(v|c', r)$, and consider the probabilities $p(v|\overline{c'_i}, r)$ only.) Note that this procedure rests on the assumption that $p(v|\overline{c}, r)$ is close to $p(v|c, r)$. (In fact, $p(v|\overline{c}, r)$ is equal to $p(v|c, r)$ when c is a leaf node.) So when finding a suitable level for the estimation of $p(\langle\text{sandwich}\rangle|\textit{eat}, \text{obj})$, for example, we first assume that $p(\textit{eat}|\overline{\langle\text{sandwich}\rangle}, \text{obj})$ is a good approximation of $p(\textit{eat}|\langle\text{sandwich}\rangle, \text{obj})$, and then apply the procedure to $p(\textit{eat}|\overline{\langle\text{sandwich}\rangle}, \text{obj})$.

A feature of the proposed generalisation procedure is that comparing probabilities of the form $p(v|C, r)$, where C is a class, is closely related to comparing ratios of probabilities of the form $p(C|v, r)/p(C|r)$ (for a given verb and argument position):

$$p(v|C, r) = \frac{p(C|v, r)}{p(C|r)} p(v|r) \tag{10}$$

Note that, for a given verb and argument position, $p(v|r)$ is constant across classes. Equation 10 is of interest because the ratio $p(C|v, r)/p(C|r)$ can be interpreted as a measure of association between the verb v and class C. This ratio is similar to pointwise mutual information (Church and Hanks, 1990), and also forms part of Resnik's association score, which will be introduced in Section 6. Thus the generalisation procedure can be thought of as one which finds "homogeneous" areas of the hierarchy, or areas consisting of classes that are associated to a similar degree with the verb (Clark and Weir, 1999).

Finally, we note that the proposed estimation method does not guarantee that the estimates form a probability distribution over the concepts in the hierarchy, and so a normalisation factor is required:

$$p_{sc}(c|v, r) = \frac{\hat{p}(v|[c, v, r], r) \frac{\hat{p}(c|r)}{\hat{p}(v|r)}}{\sum_{c' \in \mathcal{C}} \hat{p}(v|[c', v, r], r) \frac{\hat{p}(c'|r)}{\hat{p}(v|r)}} \tag{11}$$

We use p_{sc} to denote an estimate obtained using our method (since the technique finds sets of semantically similar senses, or "**S**imilarity **C**lasses"), and $[c, v, r]$ to denote the class chosen for concept c in position r of verb v; $\hat{p}$ denotes a relative frequency estimate, and $\mathcal{C}$ denotes the set of concepts in the hierarchy.

Before giving the details of the generalisation procedure, we give the relative frequency estimates of the relevant probabilities, and deal with the problem of ambiguous data. The relative frequency estimates are as follows:

$$\hat{p}(c|r) = \frac{f(c,r)}{f(r)} = \frac{\sum_{v' \in \mathcal{V}} f(c, v', r)}{\sum_{v' \in \mathcal{V}} \sum_{c' \in \mathcal{C}} f(c', v', r)} \tag{12}$$

$$\hat{p}(v|r) = \frac{f(v,r)}{f(r)} = \frac{\sum_{c' \in \mathcal{C}} f(c', v, r)}{\sum_{v' \in \mathcal{V}} \sum_{c' \in \mathcal{C}} f(c', v', r)} \tag{13}$$

$$\hat{p}(v|\overline{c'}, r) = \frac{f(\overline{c'},v,r)}{f(\overline{c'},r)} = \frac{\sum_{c'' \in \overline{c'}} f(c'', v, r)}{\sum_{v' \in \mathcal{V}} \sum_{c'' \in \overline{c'}} f(c'', v', r)} \tag{14}$$

where $f(c, v, r)$ is the number of (n, v, r) triples in the data in which n is being used to denote c, and $\mathcal{V}$ is the set of verbs in the data. The problem is that the estimates are

Table 1
Contingency table for the children of ⟨canine⟩ in the subject position of *run*.

$\overline{c_i}$	$\hat{f}(\overline{c_i}, run, \text{subj})$		$\hat{f}(\overline{c_i}, \text{subj}) - \hat{f}(\overline{c_i}, run, \text{subj})$		$\hat{f}(\overline{c_i}, \text{subj}) = \sum_{v \in \mathcal{V}} \hat{f}(\overline{c_i}, v, \text{subj})$
$\overline{\langle\text{bitch}\rangle}$	0.3	(0.5)	26.7	(26.6)	27.0
$\overline{\langle\text{dog}\rangle}$	12.8	(10.5)	620.4	(622.7)	633.2
$\overline{\langle\text{wolf}\rangle}$	0.3	(0.6)	38.7	(38.4)	39.0
$\overline{\langle\text{jackal}\rangle}$	0.0	(0.3)	20.0	(19.7)	20.0
$\overline{\langle\text{wild\_dog}\rangle}$	0.0	(0.0)	3.0	(3.0)	3.0
$\overline{\langle\text{hyena}\rangle}$	0.0	(0.2)	10.0	(9.8)	10.0
$\overline{\langle\text{fox}\rangle}$	0.0	(1.2)	72.3	(71.1)	72.3
	13.4		791.1		804.5

defined in terms of frequencies of senses, whereas the data are assumed to be in the form of (n, v, r) triples: a noun, verb and argument position. All the data used in this work have been obtained from the BNC, using the system of Briscoe and Carroll (1997), which consists of a shallow parsing component which is able to identify verbal arguments.

We take a simple approach to the problem of estimating the frequencies of senses, by distributing the count for each noun in the data evenly among all senses of the noun:

$$\hat{f}(c, v, r) = \sum_{n \in \mathsf{syn}(c)} \frac{f(n, v, r)}{|\mathsf{cn}(n)|} \tag{15}$$

where $\hat{f}(c, v, r)$ is an estimate of the number of times that concept c appears in position r of verb v, and $|\mathsf{cn}(n)|$ is the cardinality of $\mathsf{cn}(n)$. This is the approach taken by Li and Abe (1998), Ribas (1995), and McCarthy (2000).[8] Resnik (1998) explains how this apparently crude technique works surprisingly well. Alternative approaches are described in Clark and Weir (1999) (see also Clark (2001)), Abney and Light (1999) and Ciaramita and Johnson (2000).

4 Using a Chi-squared Test to Compare Probabilities

In this section we show how to test if $p(v|\overline{c'}, r)$ changes significantly by moving up a node in the hierarchy. Consider the problem of deciding if $p(run|\overline{\langle\text{canine}\rangle}, \text{subj})$ is a good approximation of $p(run|\overline{\langle\text{dog}\rangle}, \text{subj})$. (⟨canine⟩ is the parent of ⟨dog⟩ in WordNet.) To do this, the probabilities $p(run|\overline{c_i'}, \text{subj})$ are compared using a chi-squared test, where the c_i' are the children of ⟨canine⟩. In this case, the null hypothesis of the test is that the probabilities $p(run|\overline{c_i}, \text{subj})$ are the same for each child c_i. By judging the strength of the evidence against the null hypothesis, it can be determined how similar the true probabilities are likely to be. If the test indicates that the probabilities are sufficiently unlikely to be the same, then the null hypothesis is rejected, and the conclusion is that $p(run|\overline{\langle\text{canine}\rangle}, \text{subj})$ is not a good approximation of $p(run|\overline{\langle\text{dog}\rangle}, \text{subj})$.

8 Resnik takes a similar approach but divides the count evenly among the noun's senses *and* all the hypernyms of those senses.

Table 2
Contingency table for the children of ⟨liquid⟩ in the object position of *drink*.

$\overline{c_i}$	$\hat{f}(\overline{c_i}, \textit{drink}, \text{obj})$		$\hat{f}(\overline{c_i}, \text{obj}) - \hat{f}(\overline{c_i}, \textit{drink}, \text{obj})$		$\hat{f}(\overline{c_i}, \text{obj}) = \sum_{v \in \mathcal{V}} \hat{f}(\overline{c_i}, v, \text{obj})$
$\overline{\langle\text{beverage}\rangle}$	261.0	(238.7)	2,367.7	(2,390.0)	2,628.7
$\overline{\langle\text{supernatant}\rangle}$	0.0	(0.1)	1.0	(0.9)	1.0
$\overline{\langle\text{alcohol}\rangle}$	11.5	(9.4)	92.0	(94.1)	103.5
$\overline{\langle\text{ammonia}\rangle}$	0.0	(0.8)	8.5	(7.7)	8.5
$\overline{\langle\text{antifreeze}\rangle}$	0.0	(0.1)	1.0	(0.9)	1.0
$\overline{\langle\text{distillate}\rangle}$	0.0	(0.5)	6.0	(5.5)	6.0
$\overline{\langle\text{water}\rangle}$	12.0	(31.6)	335.7	(316.1)	347.7
$\overline{\langle\text{ink}\rangle}$	0.0	(2.9)	32.0	(29.1)	32.0
$\overline{\langle\text{liquor}\rangle}$	0.7	(1.1)	11.6	(11.2)	12.3
	285.2		2,855.5		3,140.7

An example contingency table, based on counts obtained from a subset of the BNC using the system of Briscoe and Carroll, is given in Table 1. (Recall that the frequencies are estimated by distributing the count for a noun equally among the noun's senses; this explains the fractional counts.) One column contains estimates of counts arising from concepts in $\overline{c_i}$ appearing in the subject position of the verb *run*: $\hat{f}(\overline{c_i}, \textit{run}, \text{subj})$. A second column contains estimates of counts arising from concepts in $\overline{c_i}$ appearing in the subject position of a verb other than *run*. The figures in brackets are the expected values if the null hypothesis is true.

There is a choice of which statistic to use in conjunction with the test. The usual statistic encountered in text books is the Pearson chi-squared statistic, denoted X^2:

$$X^2 = \sum_{i,j} \frac{(o_{ij} - e_{ij})^2}{e_{ij}} \tag{16}$$

where o_{ij} is the observed value for the cell in row i and column j, and e_{ij} is the corresponding expected value. An alternative statistic is the log-likelihood chi-squared statistic, denoted G^2:[9]

$$G^2 = 2 \sum_{i,j} o_{ij} \log_e \frac{o_{ij}}{e_{ij}} \tag{17}$$

The two statistics have similar values when the counts in the table are large (Agresti, 1996). However, the statistics behave differently when the contingency table contains low counts, and, since corpus data is likely to lead to some low counts, the question of which statistic to use is an important one. Dunning (1993) argues for the use of G^2 rather than X^2, based on an analysis of the sampling distributions of G^2 and X^2, and results obtained when using the statistics to acquire highly associated bigrams. We consider

9 An alternative formula for G^2 is given in Dunning (1993), but the two are equivalent.

Dunning's analysis at the end of this section, and the question of whether to use G^2 or X^2 will be discussed further there. For now, we continue with the discussion of how the chi-squared test is used in the generalisation procedure.

For Table 1, the value of G^2 is 3.8 and the value of X^2 is 2.5. Assuming a level of significance of $\alpha = 0.05$, the critical value is 12.6 (for 6 degrees of freedom). Thus, for this α value, the null hypothesis would not be rejected for either statistic, and the conclusion would be that there is no reason to suppose $p(run|\overline{\langle\text{canine}\rangle}, \text{subj})$ is not a reasonable approximation of $p(run|\overline{\langle\text{dog}\rangle}, \text{subj})$.

As a further example, Table 2 gives counts for the children of $\langle$liquid$\rangle$ in the object position of *drink*. Again, the counts have been obtained from a subset of the BNC using the system of Briscoe and Carroll. Not all the sets dominated by the children of $\langle$liquid$\rangle$ are shown, as some, such as $\overline{\langle\text{sheep\_dip}\rangle}$, never appear in the object position of a verb in the data. This example is designed to show a case where the null hypothesis is rejected. The value of G^2 for this table is 29.0, and the value of X^2 is 21.2. So for G^2, even if an α value as low as 0.0005 were being used (for which the critical value is 27.9 for 8 degrees of freedom), the null hypothesis would still be rejected. For X^2, the null hypothesis is rejected for α values greater than 0.005. This seems reasonable, since the probabilities associated with the children of $\langle$liquid$\rangle$ and the object position of *drink* would be expected to show a lot of variation across the children.

A key question is how to select the value for α. One solution is to treat α as a parameter and set it empirically, by taking a held-out test set and choosing the value of α that maximises performance on the relevant task. For example, Clark and Weir (2000) describes a PP-attachment algorithm that uses probability estimates obtained using the WordNet method described here. To set the value of α, the performance of the algorithm on a development set could be compared across different values of α, and the value that leads to the best performance could be chosen. Note that this approach sets no constraints on the value of α: the value could be as high as 0.995, or as low as 0.0005, depending on the particular application.

There may be cases where the conditions for the appropriate application of a chi-squared test are not met. One condition that is likely to be violated is the requirement that expected values in the contingency table should not be too small. (A rule of thumb often found in text books is that the expected values should be greater than 5.) One response to this problem is to apply some kind of thresholding, and either ignore counts below the threshold, or only apply the test to tables that do not contain low counts. Ribas (1995), Li and Abe (1998), McCarthy (2000) and Wagner (2000) all use some kind of thresholding when dealing with counts in the hierarchy (although not in the context of a chi-squared test). Another approach would be to use Fisher's exact test (Agresti, 1996; Pedersen, 1996), which can be applied to tables regardless of the size of the counts. The main problem with this test is that it is computationally expensive, especially for large contingency tables.

What we have found in practice is that applying the chi-squared test to tables dominated by low counts tends to produce an insignificant result, and the null hypothesis is not rejected. The consequences of this for the generalisation procedure are that low count tables tend to result in the procedure moving up to the next node in the hierarchy. But given that the purpose of the generalisation is to overcome the sparse data problem, this behaviour is desirable, and therefore we do not modify the test for tables with low counts.

The final issue to consider is which chi-squared statistic to use. Dunning (1993) argues for the use of G^2 rather than X^2, based on the claim that the sampling distribution of G^2 approaches the true chi-squared distribution quicker than the sampling distribu-

Algorithm $\mathsf{top}(c, v, r)$:
$\mathsf{top} \leftarrow \mathsf{c}$
$\mathsf{sig\_result} \leftarrow \mathsf{false}$
comment $\mathsf{parent}_{\mathsf{min}}$ gives lowest G^2 value, G^2_{min}
while not sig_result & top $\neq \langle\text{root}\rangle$ **do**
    $G^2_{\mathsf{min}} \leftarrow \infty$
    for all parents of top **do**
        calculate G^2 for sets dominated by children of parent
        if $G^2 < G^2_{\mathsf{min}}$
          then $G^2_{\mathsf{min}} \leftarrow G^2$
            $\mathsf{parent}_{\mathsf{min}} \leftarrow \mathsf{parent}$
    end
    if chi-squared test for $\mathsf{parent}_{\mathsf{min}}$ is significant
      then $\mathsf{sig\_result} \leftarrow \mathsf{true}$
    else move up to next node: $\mathsf{top} \leftarrow \mathsf{parent}_{\mathsf{min}}$
end
return top

Figure 1
An algorithm for determining $\mathsf{top}(c, v, r)$.

tion of X^2. However, Agresti (1996, p.34) makes the opposite claim:

> The sampling distributions of X^2 and G^2 get closer to chi-squared as the sample size n increases . . . The convergence is quicker for X^2 than G^2.

In addition, Pedersen (2001) questions whether one statistic should be preferred over the other for the bigram acquisition task, and cites Cressie and Read (1984) who argue that there are some cases where the Pearson statistic is more reliable than the log-likelihood statistic. Finally, the results of the pseudo disambiguation experiments presented in Section 7 are at least as good, if not better, when using X^2 rather than G^2, and so we conclude that the question of which statistic to use should be answered on a per-application basis.

5 The Generalisation Procedure

The procedure for finding a suitable class, $\overline{c'}$, to generalise concept c in position r of verb v works as follows. (We refer to $\overline{c'}$ as the "similarity-class" of c with respect to v and r, and the hypernym c' as $\mathsf{top}(c, v, r)$, since the chosen hypernym sits at the "top" of the similarity class.) Initially, concept c is assigned to a variable top. Then, by working up the hierarchy, successive hypernyms of c are assigned to top, which continues until the probabilities associated with the sets of concepts dominated by top and the siblings of top are significantly different. Once a node is reached which results in a significant result for the chi-squared test, the procedure stops, and top is returned as $\mathsf{top}(c, v, r)$. In cases where a concept has more than one parent, the parent is chosen which results in the lowest value of the chi-squared statistic, as this indicates the probabilities are more similar. The set $\overline{\mathsf{top}(c, v, r)}$ is the similarity-class of c for verb v and position r. Figure 1 gives an algorithm for determining $\mathsf{top}(c, v, r)$.

Figure 2 gives an example of the procedure at work. Here, $\mathsf{top}(\langle\text{soup}\rangle, \textit{stir}, \text{obj})$ is be-

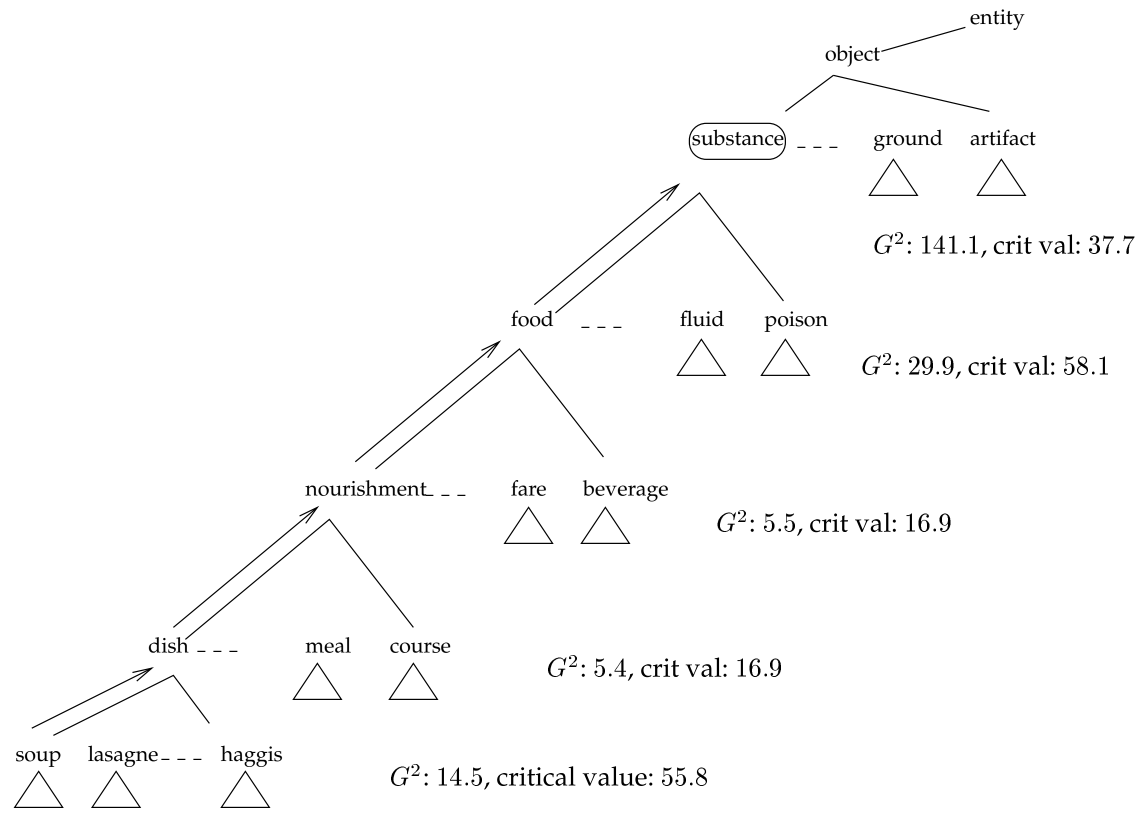

Figure 2
An example generalisation: determining top($\langle$soup$\rangle$, *stir*, obj).

ing determined. The example is based on data from a subset of the BNC, with 303 cases of an argument in the object position of *stir*. The G^2 statistic is used, together with an α value of 0.05. Initially, top is set to $\langle$soup$\rangle$, and the probabilities corresponding to the children of $\langle$dish$\rangle$ are compared: $p(stir|\overline{\langle \text{soup} \rangle}, \text{obj})$, $p(stir|\overline{\langle \text{lasagne} \rangle}, \text{obj})$, $p(stir|\overline{\langle \text{haggis} \rangle}, \text{obj})$ and so on for the rest of the children. The chi-squared test results in a G^2 value of 14.5, compared to a critical value of 55.8. Since G^2 is less than the critical value, the procedure moves up to the next node. This continues until a significant result is obtained, which first occurs at $\langle$substance$\rangle$ when comparing the children of $\langle$object$\rangle$. Thus $\langle$substance$\rangle$ is the chosen level of generalisation.

Now we show how the chosen level of generalisation varies with α, and how it varies with the size of the data set. A note of clarification is required before presenting the results. In related work on acquiring selectional preferences (Ribas, 1995; McCarthy, 1997; Li and Abe, 1998; Wagner, 2000), the level of generalisation is often determined for a small number of hand-picked verbs, and the result compared with the researcher's intuition about the most appropriate level for representing a selectional preference. According to this approach, if $\overline{\langle \text{sandwich} \rangle}$ were chosen to represent $\langle$hotdog$\rangle$ in the object position of *eat*, this might be considered an under-generalisation, since $\overline{\langle \text{food} \rangle}$ might be considered more appropriate. For this work we argue that such an evaluation is not appropriate; since the purpose of this work is probability estimation, the most appropriate level is the one that leads to the most accurate estimate, and this may or may not agree

Table 3
Example levels of generalisation for different values of α.
The selected level is shown in upper case.

(c, v, r), $f(v, r)$	α	
$(\langle$coffee$\rangle$, *drink*, obj$)$	0.0005	⟨coffee⟩⟨BEVERAGE⟩⟨food⟩ ... ⟨object⟩⟨entity⟩
	0.05	⟨coffee⟩⟨BEVERAGE⟩⟨food⟩ ... ⟨object⟩⟨entity⟩
$f($*drink*, obj$) = 849$	0.5	⟨coffee⟩⟨BEVERAGE⟩⟨food⟩ ... ⟨object⟩⟨entity⟩
	0.995	⟨coffee⟩⟨BEVERAGE⟩⟨food⟩ ... ⟨object⟩⟨entity⟩
$(\langle$hotdog$\rangle$, *eat*, obj$)$	0.0005	⟨hotdog⟩⟨sandwich⟩⟨snack_food⟩⟨DISH⟩ ... ⟨food⟩ ... ⟨entity⟩
	0.05	⟨hotdog⟩⟨sandwich⟩⟨snack_food⟩⟨DISH⟩ ... ⟨food⟩ ... ⟨entity⟩
$f($*eat*, obj$) = 1,703$	0.5	⟨hotdog⟩⟨sandwich⟩⟨snack_food⟩⟨DISH⟩ ... ⟨food⟩ ... ⟨entity⟩
	0.995	⟨hotdog⟩⟨SANDWICH⟩⟨snack_food⟩⟨dish⟩ ... ⟨food⟩ ... ⟨entity⟩
$(\langle$Socrates$\rangle$, *kiss*, obj$)$	0.0005	⟨Socrates⟩ ... ⟨person⟩⟨life_form⟩⟨CAUSAL_AGENT⟩⟨entity⟩
	0.05	⟨Socrates⟩ ... ⟨person⟩⟨life_form⟩⟨CAUSAL_AGENT⟩⟨entity⟩
$f($*kiss*, obj$) = 345$	0.5	⟨Socrates⟩ ... ⟨person⟩⟨life_form⟩⟨CAUSAL_AGENT⟩⟨entity⟩
	0.995	⟨Socrates⟩ ... ⟨PERSON⟩⟨life_form⟩⟨causal_agent⟩⟨entity⟩
$(\langle$dream$\rangle$, *remember*, obj$)$	0.0005	⟨dream⟩ ... ⟨preoccupation⟩⟨cognitive_state⟩⟨STATE⟩
	0.05	⟨dream⟩ ... ⟨preoccupation⟩⟨cognitive_state⟩⟨STATE⟩
$f($*remember*, obj$) = 1,982$	0.5	⟨dream⟩ ... ⟨preoccupation⟩⟨COGNITIVE_STATE⟩⟨state⟩
	0.995	⟨dream⟩ ... ⟨PREOCCUPATION⟩⟨cognitive_state⟩⟨state⟩
$(\langle$man$\rangle$, *see*, obj$)$	0.0005	⟨man⟩ ... ⟨mammal⟩ ... ⟨ANIMAL⟩⟨life_form⟩⟨entity⟩
	0.05	⟨man⟩ ... ⟨MAMMAL⟩ ... ⟨animal⟩⟨life_form⟩⟨entity⟩
$f($*see*, obj$) = 16,757$	0.5	⟨man⟩ ... ⟨MAMMAL⟩ ... ⟨animal⟩⟨life_form⟩⟨entity⟩
	0.995	⟨MAN⟩ ... ⟨mammal⟩ ... ⟨animal⟩⟨life_form⟩⟨entity⟩
$(\langle$belief$\rangle$, *abandon*, obj$)$	0.0005	⟨belief⟩⟨mental_object⟩⟨cognition⟩⟨PSYCH_FEATURE⟩
	0.05	⟨belief⟩⟨MENTAL_OBJECT⟩⟨cognition⟩⟨psychological_feature⟩
$f($*abandon*, obj$) = 673$	0.5	⟨BELIEF⟩⟨mental_object⟩⟨cognition⟩⟨psychological_feature⟩
	0.995	⟨BELIEF⟩⟨mental_object⟩⟨cognition⟩⟨psychological_feature⟩
$(\langle$nightmare$\rangle$, *have*, obj$)$	0.0005	⟨nightmare⟩⟨dreaming⟩⟨IMAGINATION⟩ ... ⟨psych_feature⟩
	0.05	⟨nightmare⟩⟨dreaming⟩⟨IMAGINATION⟩ ... ⟨psych_feature⟩
$f($*have*, obj$) = 93,683$	0.5	⟨nightmare⟩⟨DREAMING⟩⟨imagination⟩ ... ⟨psych_feature⟩
	0.995	⟨nightmare⟩⟨DREAMING⟩⟨imagination⟩ ... ⟨psych_feature⟩

with intuition. Furthermore, we show in Section 7 that to generalise unnecessarily can be harmful for some tasks: if we already have lots of data regarding ⟨sandwich⟩, why generalise any higher? Thus the purpose of this section is not to show that the acquired levels are "correct," but simply to show how the levels vary with α and the sample size.

To show how the level of generalisation varies with changes in α, $\texttt{top}(c, v, \text{obj})$ was determined for a number of hand-picked (c, v, obj) triples over a range of values for α. The triples were chosen to give a range of strongly and weakly selecting verbs and a range of verb frequencies. The data were again extracted from a subset of the BNC using the system of Briscoe and Carroll (1997), and the G^2 statistic was used in the chi-squared test. The results are shown in Table 3. The number of times the verb occurred with some object is also given in the table.

Table 4
The extent of generalisation for different values of α and sample sizes.

α	100%	50%	10%	1%
0.0005	3.3	3.9	5.0	5.6
0.05	2.8	3.5	4.6	5.6
0.5	2.1	2.9	4.1	5.4
0.995	1.2	1.5	2.6	3.9

The results suggest that the generalisation level becomes more specific as α increases. This is to be expected, since, given a contingency table chosen at random, a higher value of α is more likely to lead to a significant result than a lower value of α. We also see that, for some cases, the value of α has little effect on the level. We would expect there to be less change in the level of generalisation for strongly selecting verbs, such as *drink* and *eat*, and a greater range of levels for weakly selecting verbs such as *see*. This is because any significant difference in probabilities is likely to be more marked for a strongly selecting verb, and likely to be significant over a wider range of α values. The table only provides anecdotal evidence, but provides some support to this argument.

To investigate more generally how the level of generalisation varies with changes in α, and also with changes in sample size, we took $6,000$ (c, v, obj) triples and calculated the difference in depth between c and $\mathtt{top}(c, v, r)$ for each triple. The $6,000$ triples were taken from the first experimental test set described in Section 7, and the training data from this experiment were used to provide the counts. (The test set contains nouns, rather than noun senses, and so the sense of the noun which is most probable given the verb and object slot was used.) An average difference in depth was then calculated. To give an example of how the difference in depth was calculated, suppose $\langle$dog$\rangle$ generalised to $\langle$placental_mammal$\rangle$ via $\langle$canine$\rangle$ and $\langle$carnivore$\rangle$; in this case the difference would be 3.

The results for various levels of α and different sample sizes are shown in Table 4. The figures in each $x\%$ column arise from using the contingency tables based on the complete training data, but with each count in the table multiplied by $x\%$. Thus the 50% column is based on contingency tables where each original count is multiplied by 50%, which is equivalent to using a sample one-half the size of the original training set. Reading across a row shows how the generalisation varies with sample size, and reading down a column shows how it varies with α. The results show clearly that the extent of generalisation decreases with an increase in the value of α, supporting the trend observed in Table 3. The results also show that the extent of generalisation increases with a decrease in sample size. Again, this is to be expected, since any difference in probability estimates is less likely to be significant for tables with low counts.

6 Alternative Class-Based Estimation Methods

The approaches used for comparison are those of Resnik (Resnik, 1993; Resnik, 1998), subsequently developed by Ribas (1995), and Li and Abe (1998), which has been adopted by McCarthy (2000). These have been chosen because they directly address the question of how to find a suitable level of generalisation in WordNet.

The first alternative uses the "association score", which is a measure of how well a set of concepts, C, satisfies the selectional preferences of a verb, v, for argument position, r:[10]

$$\mathrm{A}(C, v, r) = p(C|v, r) \log_2 \frac{p(C|v, r)}{p(C|r)} \tag{18}$$

An estimate of the association score, $\hat{\mathrm{A}}(C, v, r)$, can be obtained using relative frequency estimates of the probabilities. The key question is how to determine a suitable level of generalisation for concept c, or, alternatively, how to find a suitable class to represent concept c (assuming the choice is from those classes that contain all concepts dominated by some hypernym of c). Resnik's solution to this problem (which he neatly refers to as the "vertical ambiguity" problem) is to choose the class that maximises the association score.

It is not clear that the class with the highest association score is always the most appropriate level of generalisation. For example, this approach does not always generalise appropriately for arguments that are *negatively* associated with some verb. To see why, consider the problem of deciding how well the concept ⟨location⟩ satisfies the preferences of the verb *eat* for its object. Since locations are not the kinds of things that are typically eaten, a suitable level of generalisation would correspond to a class that has a low association score with respect to *eat*. However, ⟨location⟩ is a kind of ⟨entity⟩ in WordNet,[11] and choosing the class with the highest association score is likely to produce $\overline{\langle\text{entity}\rangle}$ as the chosen class. This is a problem, because the association score of $\overline{\langle\text{entity}\rangle}$ with respect to *eat* may be too high to reflect the fact that ⟨location⟩ is a very unlikely object of the verb.

Note that the solution to the vertical ambiguity problem presented in the previous sections is able to generalise appropriately in such cases. Continuing with the *eat* ⟨location⟩ example, our generalisation procedure is unlikely to get as high as ⟨entity⟩ (assuming a reasonable number of examples of *eat* in the training data), since the probabilities corresponding to the daughters of ⟨entity⟩ are likely to be very different with respect to the object position of *eat*.

The second alternative uses the Minimum Description Length principle (MDL). Li and Abe use MDL to select a set of classes from a hierarchy, together with their associated probabilities, to represent the selectional preferences of a verb. The preferences and class-based probabilities are then used to estimate probabilities of the form $p(n|v, r)$, where n is a noun, v is a verb and r is an argument slot.

Li and Abe's application of MDL requires the hierarchy to be in the form of a thesaurus, where each leaf node represents a noun, and internal nodes represent the class of nouns that the node dominates. The hierarchy is also assumed to be in the form of a tree. The class-based models consist of a partition of the set of nouns (leaf nodes) and a probability associated with each class in the partition. The probabilities are the conditional probabilities of each class, given the relevant verb and argument position. Li and Abe refer to such a partition as a "cut", and the cut together with the probabilities, a "tree cut model". The probabilities of the classes in a cut, Γ, satisfy the following constraint:

$$\sum_{C \in \Gamma} p(C|v, r) = 1 \tag{19}$$

10 The definition used here is that given by Ribas (1995).
11 For example, the hypernyms of the concept ⟨Dallas⟩ are as follows: ⟨city⟩, ⟨municipality⟩, ⟨urban_area⟩, ⟨geographical_area⟩, ⟨region⟩, ⟨location⟩, ⟨object⟩, ⟨entity⟩.

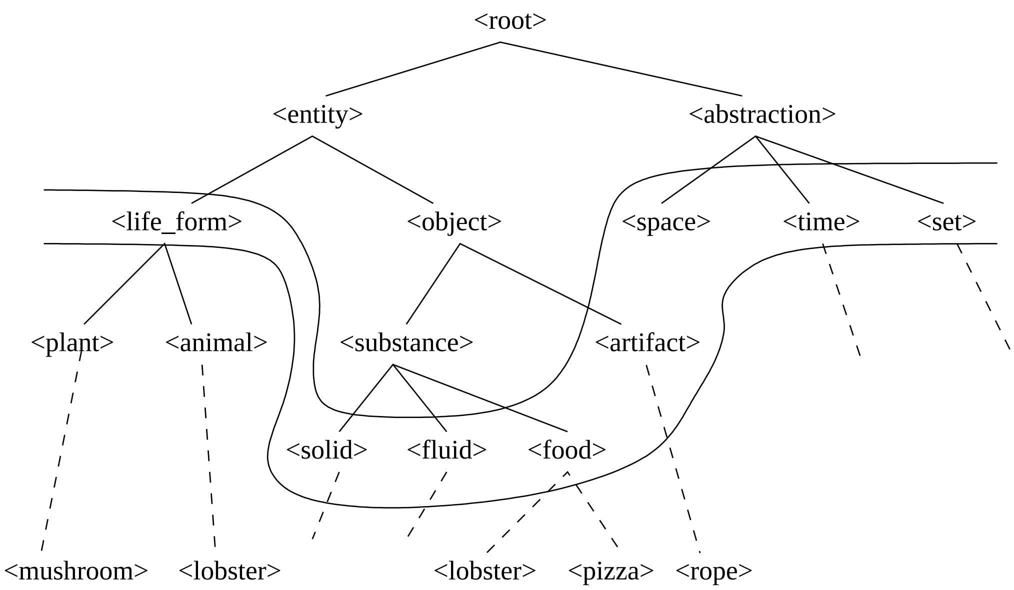

Figure 3
Possible cut returned by MDL.

In order to determine the probability of a noun, the probability of a class is assumed to be distributed uniformly among the members of that class:

$$p(n|v,r) = \frac{1}{|C|}\, p(C|v,r) \quad \text{for all } \; n \in C \tag{20}$$

Since WordNet is a hierarchy with noun senses at the nodes, rather than nouns, Li and Abe deal with the issue of word sense ambiguity using the method described in Section 3, by dividing the count for a noun equally among the concepts whose synsets contain the noun. Also, since WordNet is a DAG, Li and Abe turn WordNet into a tree by copying each subgraph with multiple parents. And in order that each noun in the data appears (in a synset) at a leaf node, Li and Abe remove those parts of the hierarchy dominated by a noun in the data (but only for that instance of WordNet corresponding to the relevant verb).

An example cut showing part of the WordNet hierarchy is shown in Figure 3 (based on an example from Li and Abe (1998); the dashed lines indicate parts of the hierarchy that are not shown in the diagram). This is a possible cut for the object position of the verb *eat*, and the cut consists of the following classes: $\overline{\langle \text{life\_form} \rangle}$, $\overline{\langle \text{solid} \rangle}$, $\overline{\langle \text{fluid} \rangle}$, $\overline{\langle \text{food} \rangle}$, $\overline{\langle \text{artifact} \rangle}$, $\overline{\langle \text{space} \rangle}$, $\overline{\langle \text{time} \rangle}$, $\overline{\langle \text{set} \rangle}$. (The particular choice of classes for the cut in this example is not too important; the example is designed to show how probabilities of senses are estimated from class probabilities.) Since the class in the cut containing $\langle \text{pizza} \rangle$ is $\overline{\langle \text{food} \rangle}$, the probability $p(\langle \text{pizza} \rangle | \mathit{eat}, \text{obj})$ would be estimated as $p(\overline{\langle \text{food} \rangle} | \mathit{eat}, \text{obj}) / |\overline{\langle \text{food} \rangle}|$. Similarly, since the class in the cut containing $\langle \text{mushroom} \rangle$ is $\overline{\langle \text{life\_form} \rangle}$, the probability $p(\langle \text{mushroom} \rangle | \mathit{eat}, \text{obj})$ would be estimated as $p(\overline{\langle \text{life\_form} \rangle} | \mathit{eat}, \text{obj}) / |\overline{\langle \text{life\_form} \rangle}|$.

The uniform distribution assumption (20) means that cuts close to the root of the hierarchy result in a greater smoothing of the probability estimates than cuts near to the leaves. Thus there is a trade-off between choosing a model that has a cut near the leaves, which is likely to over-fit the data, and a more general (simple) model near the root, which is likely to under-fit the data. MDL looks ideally suited to the task of model selection, since it is designed to deal with precisely this trade-off. The simplicity of a model is measured using the *model description length*, which is an information-theoretic term and denotes the number of bits required to encode the model. The fit to the data is measured using the *data description length*, which is the number of bits required to encode the data (relative to the model). The overall description length is the sum of the model description length and the data description length, and the MDL principle is to select the model with the shortest description length.

We used McCarthy's (2000) implementation of MDL. In order that every noun is represented at a leaf node, McCarthy does not remove parts of the hierarchy, as Li and Abe do, but creates new leaf nodes for each synset at an internal node. However, unlike Li and Abe, McCarthy does not transform WordNet into a tree, which is strictly required for Li and Abe's application of MDL. This did create a problem, in that many of the cuts returned by MDL were over-generalising at the $\langle$entity$\rangle$ node. The reason is that $\langle$person$\rangle$, which is close to $\langle$entity$\rangle$, and dominated by $\langle$entity$\rangle$, has two parents: $\langle$life_form$\rangle$ and $\langle$causal_agent$\rangle$. This DAG-like property was responsible for the over-generalisation, and so we removed the link between $\langle$person$\rangle$ and $\langle$causal_agent$\rangle$. This appeared to solve the problem, and the results presented later for the average degree of generalisation do not show an over-generalisation compared with those given in Li and Abe (1998).

7 Pseudo Disambiguation Experiments

The task we used to compare the class-based estimation techniques is a decision task previously used by Pereira, Tishby, and Lee (1993) and Rooth et al. (1999). The task is to decide which of two verbs, v and v', is more likely to take a given noun, n, as an object. The test and training data were obtained as follows. A number of verb direct object pairs were extracted from a subset of the BNC, using the system of Briscoe and Carroll. All those pairs containing a noun not in WordNet were removed, and each verb and argument was lemmatised. This resulted in a data set of around 1.3 million (v, n) pairs.

To form a test set, $3,000$ of these pairs were randomly selected, such that each selected pair contained a fairly frequent verb. (Following Pereira et al., only those verbs that occurred between 500 and $5,000$ times in the data were considered.) Each instance of a selected pair was then deleted from the data. This was to ensure that the test data were unseen. The remaining pairs formed the training data. To complete the test set, a further fairly frequent verb, v', was randomly chosen for each (v, n) pair. The random choice was made according to the verb's frequency in the original data set, subject to the condition that the pair (v', n) did not occur in the training data. Given the set of (v, n, v') triples, the task is to decide whether (v, n) or (v', n) is the correct pair.[12]

We acknowledge that the task is somewhat artificial, but pseudo-disambiguation tasks of this kind are becoming popular in statistical NLP because of the ease with which training and test data can be created. We also feel that the pseudo-disambiguation task is useful for evaluating the different estimation methods, since it directly addresses the question of how likely a predicate is to take a given noun as an argument. An evaluation using a PP-attachment task was attempted in Clark and Weir (2000), but the evaluation was limited by the relatively small size of the Penn Treebank.

Using our approach, the disambiguation decision for each (v, n, v') triple was made

12 We note that this procedure does not guarantee that the correct pair is more likely than the incorrect pair, because of noise in the data from the parser, and also because a highly plausible incorrect pair could be generated by chance.

Table 5
Results for the pseudo disambiguation task.
av.gen. is the average number of generalised levels; sd.gen. is the standard deviation.

Generalisation technique	% correct	av.gen.	sd.gen
Similarity-class			
$\alpha = 0.0005$	73.8	3.3	2.0
$\alpha = 0.05$	73.4	2.8	1.9
$\alpha = 0.3$	73.0	2.4	1.8
$\alpha = 0.75$	73.9	1.9	1.6
$\alpha = 0.995$	73.8	1.2	1.2
Low-class	73.6	0.9	1.0
MDL	68.3	4.1	1.9
Assoc	63.9	4.2	2.1

according to the following procedure:

if $\max_{c \in \mathsf{cn}(n)} p_{sc}(c|v, \text{obj}) > \max_{c \in \mathsf{cn}(n)} p_{sc}(c|v', \text{obj})$
then choose (v, n)
else if $\max_{c \in \mathsf{cn}(n)} p_{sc}(c|v', \text{obj}) > \max_{c \in \mathsf{cn}(n)} p_{sc}(c|v, \text{obj})$
then choose (v', n)
else choose at random

If n has more than one sense, the sense is chosen that maximises the relevant probability estimate; this explains the maximisation over $\mathsf{cn}(n)$. The probability estimates were obtained using our class-based method, and the G^2 statistic was used for the chi-squared test. This procedure was also used for the MDL alternative, but using the MDL method to estimate the probabilities.

Using the association score, the decision for each test triple was made according to the following procedure:

if $\max_{c \in \mathsf{cn}(n)} \max_{c' \in \mathsf{h}(c)} \hat{\mathrm{A}}(\overline{c'}, v, \text{obj}) > \max_{c \in \mathsf{cn}(n)} \max_{c' \in \mathsf{h}(c)} \hat{\mathrm{A}}(\overline{c'}, v', \text{obj})$
then choose (v, n)
else if $\max_{c \in \mathsf{cn}(n)} \max_{c' \in \mathsf{h}(c)} \hat{\mathrm{A}}(\overline{c'}, v', \text{obj}) > \max_{c \in \mathsf{cn}(n)} \max_{c' \in \mathsf{h}(c)} \hat{\mathrm{A}}(\overline{c'}, v, \text{obj})$
then choose (v', n)
else choose at random

We use $\mathsf{h}(c)$ to denote the set consisting of the hypernyms of c. The inner maximisation is over $\mathsf{h}(c)$, assuming c is the chosen sense of n, which corresponds to Resnik's method of choosing a set to represent c. The outer maximisation is over the senses of n, $\mathsf{cn}(n)$, which determines the sense of n by choosing the sense that maximises the association score.

Table 6
Results for the pseudo disambiguation task with 1/5th training data.
av.gen. is the average number of generalised levels; sd.gen. is the standard deviation.

Generalisation technique	% correct	av.gen.	sd.gen
Similarity-class			
$\alpha = 0.0005$	66.7	4.5	1.9
$\alpha = 0.05$	68.4	4.1	1.9
$\alpha = 0.3$	70.2	3.7	1.9
$\alpha = 0.75$	72.3	3.0	1.9
$\alpha = 0.995$	72.4	1.9	1.6
Low-class	71.9	1.1	1.1
MDL	62.9	4.7	1.9
Assoc	62.6	4.1	2.0

The first set of results is given in Table 5. Our technique is referred to as the "similarity-class" technique, and the approach using the association score is referred to as "Assoc". The results are given for a range of α values, and demonstrate clearly that the performance of similarity-class varies little with changes in α, and similarity-class outperforms both MDL and Assoc.[13]

We also give a score for our approach using a simple generalisation procedure, which we call "Low-class". The procedure is to select the first class that has a count greater than zero (relative to the verb and argument position), which is likely to return a low level of generalisation, on the whole. The results show that our generalisation technique only narrowly outperforms the simple alternative. Note that, although "Low-class" is based on a very simple generalisation method, the estimation method is still using our class-based technique, by applying Bayes' theorem and conditioning on a class, as described in Section 3; the difference is in how the class is chosen.

To investigate the results, we calculated the average number of generalised levels for each approach. The number of generalised levels for a concept c (relative to a verb v and argument position r) is the difference in depth between c and $\texttt{top}(c, v, r)$, as explained in Section 5. For each test case, the number of generalised levels for both verbs, v and v', was calculated, but only for the chosen sense of n. The results are given in the third column of Table 5, and demonstrate clearly that both MDL and Assoc are generalising to a greater extent than similarity-class. (The fourth column gives a standard deviation figure.) These results suggest that MDL and Assoc are over-generalising, at least for the purposes of this task.

To investigate why the value for α had no impact on the results, we repeated the experiment, but with 1/5th of the data. A new data set was created by taking every 5th pair of the original 1.3 million pairs. A test set of $3,000$ triples was created from this new data set, as before, but this time only verbs that occurred between 100 and $1,000$ times were considered. The results using these test and training data are given in Table 6.

These results show a variation in performance across values for α, with an optimal performance when α is around 0.75. (Of course, in practice, the value for α would need

13 The results given for similarity-class are different to those given in Clark and Weir (2001) because the probability estimates used in Clark and Weir (2001) were not normalised.

Table 7
Disambiguation results for G^2 and X^2.

α value	% correct – G^2	% correct – X^2
0.0005	73.8 (3.3)	74.1 (3.0)
0.05	73.4 (2.8)	73.8 (2.5)
0.3	73.0 (2.4)	74.1 (2.2)
0.75	73.9 (1.9)	74.3 (1.8)
0.995	73.8 (1.2)	73.3 (1.2)

to be optimised on a held-out set.) But even with this variation, similarity-class is still out-performing MDL and Assoc across the whole range of α values. Note that the α values corresponding to the lowest scores lead to a significant amount of generalisation, which provides additional evidence that MDL and Assoc are over-generalising for this task. The Low-class method scores highly for this data set also, but given that the task is one that apparently favours a low level of generalisation, the high score is not too surprising.

As a final experiment, we compared the task performance using the X^2, rather than G^2, statistic in the chi-squared test. The results are given in Table 7 for the complete data set.[14] The figures in brackets give the average number of generalised levels. The X^2 statistic is performing at least as well as G^2, and the results show that the average level of generalisation is slightly higher for G^2 than X^2. This suggests a possible explanation for the results presented here, and those in Dunning (1993), which is that the X^2 statistic provides a less conservative test when counts in the contingency table are low. (By a conservative test we mean one in which the null hypothesis is not easily rejected). A less conservative test is better suited to the pseudo disambiguation task, since this results in a lower level of generalisation, on the whole, which is good for this task. In contrast, the task that Dunning considers, the discovery of bigrams, is better served by a more conservative test.

8 Conclusion

We have presented a class-based estimation method which incorporates a procedure for finding a suitable level of generalisation in WordNet. This method has been shown to provide superior performance on a pseudo disambiguation task, compared with two alternative approaches. An analysis of the results has shown that the other approaches appear to be over-generalising, at least for this task. One of the features of the generalisation procedure is the way that α, the level of significance in the chi-squared test, is treated as a parameter. This allows some control over the extent of generalisation, which can be tailored to particular tasks. We have also shown that the task performance is at least as good when using the Pearson chi-squared statistic, as when using the log-likelihood chi-squared statistic.

There are a number of ways in which this work could be extended. One possibility would be to use all the classes dominated by the hypernyms of a concept, rather than just one, to estimate the probability of the concept. An estimate would be obtained

14 X^2 performed slightly better than G^2 using the smaller data set also.

for each hypernym, and the estimates combined in a linear interpolation. An approach similar to this is taken by Bikel (2000), in the context of statistical parsing.

There is still room for investigation of the hidden data problem when using data that has not been sense disambiguated. In this paper, a very simple approach is taken, which is to split the count for a noun evenly among the noun's senses. Abney and Light (1999) have tried a more motivated approach, using the EM algorithm, but with little success. The approach described in Clark and Weir (1999) is shown in Clark (2001) to have some impact on the pseudo disambiguation task, but only with certain values of the α parameter, and ultimately does not improve on the best performance.

Finally, an issue that has not been much addressed in the literature (except by Li and Abe (1996)) is how the accuracy of class-based estimation techniques compare when using automatically acquired classes, as opposed to the manually created classes from WordNet. The pseudo disambiguation task described here has also been used to evaluate clustering algorithms (Pereira, Tishby, and Lee, 1993; Rooth et al., 1999), but with different data, and so it is difficult to compare the results. A related issue is how the structure of WordNet affects the accuracy of the probability estimates. We have taken the structure of the hierarchy for granted, without any analysis, but it may be that an alternative design could be more conducive to probability estimation.

Acknowledgments

This work was carried out while the first author was a DPhil student at the University of Sussex, and was supported by an EPSRC studentship. We would like to thank Diana McCarthy for suggesting the pseudo disambiguation task and providing the MDL software, John Carroll for supplying the data, and Ted Briscoe, Geoff Sampson, Gerald Gazdar, Bill Keller, Ted Pedersen and the anonymous reviewers for their helpful comments. We would also like to thank Ted Briscoe for presenting an earlier version of this paper on our behalf at NAACL 2001.

References

Abney, Steven P. and Marc Light. 1999. Hiding a semantic hierarchy in a Markov model. In *Proceedings of the ACL Workshop on Unsupervised Learning in Natural Language Processing*, University of Maryland, MD.

Agirre, Eneko and David Martinez. 2001. Learning class-to-class selectional preferences. In *Proceedings of the 5th ACL Workshop on Computational Language Learning*, Toulouse, France.

Agresti, Alan. 1996. *An Introduction to Categorical Data Analysis*. Wiley.

Bikel, Daniel M. 2000. A statistical model for parsing and word-sense disambiguation. In *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, pages 155–163, Hong Kong.

Briscoe, Ted and John Carroll. 1997. Automatic extraction of subcategorisation from corpora. In *Proceedings of the Fifth ACL Conference on Applied Natural Language Processing*, pages 356–363, Washington, DC.

Church, Kenneth W. and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29.

Ciaramita, Massimiliano and Mark Johnson. 2000. Explaining away ambiguity: Learning verb selectional preference with Bayesian networks. In *Proceedings of the 18th International Conference on Computational Linguistics*, pages 187–193, Saarbrucken, Germany.

Clark, Stephen. 2001. *Class-Based Statistical Models for Lexical Knowledge Acquisition*. Ph.D. thesis, University of Sussex.

Clark, Stephen 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, University of Maryland, MD.

Clark, Stephen and David Weir. 2000. A class-based probabilistic approach to structural disambiguation. In *Proceedings of the 18th International Conference on Computational Linguistics*, pages 194–200, Saarbrucken, Germany.

Clark, Stephen and David Weir. 2001. Class-based probability estimation using a semantic hierarchy. In *Proceedings of the Second Meeting of the North American*

Chapter of the Association for Computational Linguistics, pages 95–102, Pittsburgh, PA.

Cressie, Noel A. C. and Timothy R. C. Read. 1984. Multinomial goodness of fit tests. *Journal of the Royal Statistics Society Series B*, 46:440–464.

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

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

Li, Hang and Naoki Abe. 1996. Clustering words with the MDL principle. In *Proceedings of the 16th International Conference on Computational Linguistics*, pages 4–9, Copenhagen, Denmark.

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

McCarthy, Diana. 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*, pages 52–61, Madrid, Spain.

McCarthy, Diana. 2000. Using semantic preferences to identify verbal participation in role switching. In *Proceedings of the first Conference of the North American Chapter of the Association for Computational Linguistics*, pages 256–263, Seattle, WA.

Miller, George A. 1998. Nouns in WordNet. In Christiane Fellbaum, editor, *WordNet: An Electronic Lexical Database*. The MIT Press, Cambridge, MA, chapter 1, pages 23–46.

Pedersen, Ted. 1996. Fishing for exactness. In *Proceedings of the South–Central SAS Users Group Conference*, Austin, TX.

Pedersen, Ted. 2001. A decision tree of bigrams is an accurate predictor of word sense. In *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics*, pages 79–86, Pittsburgh, PA.

Pereira, Fernando, Naftali Tishby, and Lillian Lee. 1993. Distributional clustering of English words. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, pages 183–190, Columbus, OH.

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

Resnik, Philip. 1998. Wordnet and class-based probabilities. In Christiane Fellbaum, editor, *WordNet: An Electronic Lexical Database*. The MIT Press, chapter 10, pages 239–263.

Ribas, Francesc. 1995. On learning more appropriate selectional restrictions. In *Proceedings of the 7th Conference of the European Chapter of the Association for Computational Linguistics*, pages 112–118, Dublin, Ireland.

Rooth, Mats, Stefan Riezler, Detlef Prescher, Glenn Carroll, and Franz Beil. 1999. Inducing a semantically annotated lexicon via EM-based clustering. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 104–111, University of Maryland, MD.

Wagner, Andreas. 2000. Enriching a lexical semantic net with selectional preferences by means of statistical corpus analysis. In *Proceedings of the ECAI-2000 Workshop on Ontology Learning*, Berlin, Germany.