Class-Based Statistical Models for Lexical Knowledge Acquisition

Stephen Clark

CSRP 539

September, 2001

ISSN 1350–3162

UNIVERSITY OF

Cognitive Science
Research Papers

This thesis is about the automatic acquisition of a particular kind of lexical knowledge namely the knowledge of which noun senses can ll the argument slots of predicates Knowledge of this kind is closely related to the classical notion of selectional restrictions Katz and Fodor 4 and selectional preferences Wilks ; Resnik a However there is a difference in that selectional restrictions and preferences are usually expressed as constraints on the semantic class of an argument; a much used example is that the verb *rh* constrains its object to be a kind of liquid or the verb *rh* strongly prefers a kind of liquid The purpose of this thesis is not to

This basic approach can be applied to other problems such as anaphora resolution and word sense disambiguation Consider the problem of determining the referent of *i* in the following sentence taken from Wilks

I bought the wine sat on a rock and drank it

To determine the correct referent we can use the fact that the correct sense of *n* is more likely to be an object of *r n* than the correct sense of *ro*

Resnik a argues that the constraints a predicate places on its arguments are not Boolean constraints as in the classical account of selectional restrictions Katz and Fodor 4 but that the constraints are satis ed to a certain degree Resnik cites McCawley and Fodor as earlier critics of Katz and Fodor s theory We follow Resnik in modelling the constraints as graded preferences and in line with other recent work in this area Ribas b; Li and Abe ; McCarthy 2 ; Wagner 2 probabilities are used to encode the preferences An impor tant question is whether the preference measure should de ne a probability distribution over the possible arguments of a predicate

Resnik s measure of selectional preference which he calls selectional association is de ned in terms of probabilities but the measure does not de ne a probability distribution over the pos sible arguments of a predicate; the values for selectional association need not lie between zero and one and do not sum to one over the possible arguments This is also true of a number of related measures in the literature such as the chi squared statistic Kilgarriff likelihood ratio statistics Dunning and mutual information Church and Hanks Aside from the question of whether these measures are appropriate for use in corpus based linguistics Dunning they all suffer from a limitation

The limitation arises when determining the semantic plausibility of a complex linguistic event such as a parse tree In order to do parse selection one can measure the overall extent

This chapter is divided into two sections; one section describes work from those areas of lexical acquisition that are of particular relevance to this thesis and the other section describes previous approaches to structural disambiguation and parse selection These areas of application are considered because the problems of structural disambiguation and parse selection are dealt with in Chapters and

The knowledge acquisition section focuses on selectional preferences describing in detail those approaches that have used WordNet and showing how they relate to the class based estimation method described in Chapter We also describe some approaches to automatic clustering which is an important alternative to using a man made hierarchy for generalisation and also collocation extraction which has used statistics that are used in Chapters and 4 Finally a number of smoothing techniques for probability estimation are described; this work is relevant because the class based estimation method described in Chapter can be thought of as performing a kind of smoothing

The applications section focuses on those approaches to structural disambiguation and parse selection that have used knowledge similar to lexical sense preferences; this includes much of the recent work on resolving PP attachment ambiguities and statistical parsing where there has been a move towards probability models based on lexical dependencies

The role of the lexicon has taken on increasing importance in recent years both from a theoretical and a computational perspective O e 4 l B e 2 d 4 bl o s 4 e 2a 2 f 2 T
Thaeapp

arguments but rather has a *pr rr* kind of argument However Wilks distanced himself from a probabilistic treatment of preferences it is still the case that an individual preference is either satis ed or it is not as with selectional restrictions The difference is that an interpretation of a sentence can be preferred even if individual preferences are violated as long as there is no alternative interpretation with less violations

Resnik a took the notion of preference one step further by suggesting that preference should be measured on a continuous scale Resnik uses the following list of examples which orig inally appeared in Drange to demonstrate that the preferences of *o*

sn s o o s on pr r n

The parts of Resnik s work a b a b that are most relevant for this thesis are his solutions to the following questions

How can a probability distribution over the WordNet hierarchy be de ned [2]

2 How can we measure the extent to which an argument satis es the preferences of a predi cate

Each question will be dealt with in turn

Resnik de nes his probability model in terms of classes where *ss* has the interpretation given above Let $C = \{ c , c_2, \ldots , c \}$ be the set of classes in WordNet where is the number of concepts so that each concept has a corresponding class Resnik places the following constraints on any probability distribution over C

$$\text{if } c_j \text{ is a kind of } c \text{ then } p(c) \geq p(c_j) \qquad 2$$

$$\sum p(c_j) = \qquad 2\,4$$

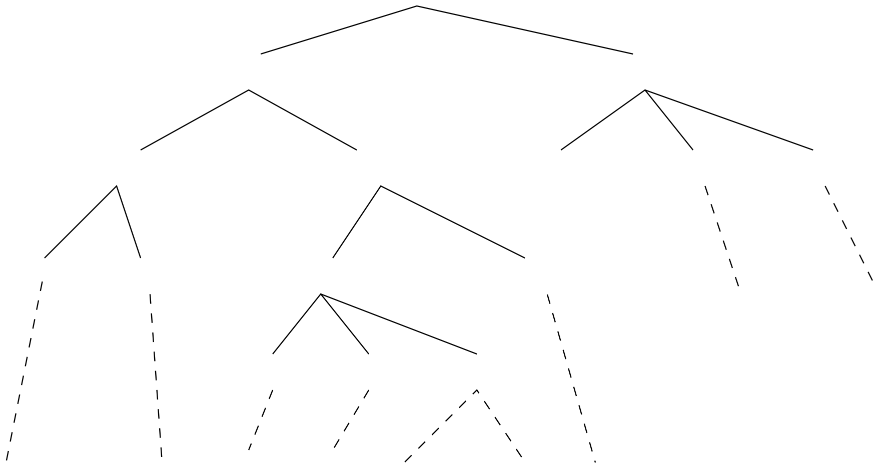
Equation 2 agrees with the intuition that the probability of a class increases with the level of abstraction Although note that the probability corresponding to a node in the hierarchy is not de ned in terms of the sum of the probabilities of the children Equation 2 4 is required by Resnik because he de nes a random variable ranging over all the classes and de nes information theoretic functions of that random variable such as entropy

Resnik s aim is to model the fact that some verbs select more strongly for their arguments than others For example selects more strongly for its direct object than *n* Resnik s approach is based on the fact that for strongly selecting verbs the probability of a class conditional on the verb $p(c|)$ is likely to differ largely from the unconditional probability $p(c)$ From an information theoretic perspective a strongly selecting verb provides more information about the

C p r _ r ous or

A dif culty with using selectional association in an application is that the arguments are likely to be nouns rather than classes and so an appropriate class has to be chosen for the noun This problem has two dimensions since a noun can have more than one sense but can also be repre

BEVERAGE FOOD LIQUID FLUID ... ENTITY Each of these classes would receive a count of /2 for each instance of in the data Note that this method of class estimation is unusual among the work in this area and is motivated by the desire to de ne a probability distribution over the set of all classes The other work described here does not



relative to the entire data size and the number of words it generalizes them into a class When the differences are especially noticeable relative to the entire data size and the number of the words on the other hand it stops generalization at that level

As we shall see a similar approach to generalization is taken in this thesis but not using MDL

One of the problems with this generalization approach is that it is based on frequencies which considering The rst modi cation is based on the following observation that removing parts of the hierarchy based on the nouns that occur in the data can result in large parts being excised For example if *n* appeared in the data a large proportion of the complete hierarchy would be removed namely that part of the hierarchy dominated by ⟨`entity`⟩ McCarthy s alternative solution is to create new leaf nodes for each internal node in the hierarchy; for example the synset for the concept ⟨`entity`⟩ would be represented at a new leaf node having the internal ⟨`entity`⟩ node as a parent This modi cation results in all the nouns in the hierarchy being represented at leaf nodes Counts for nouns are distributed initially at leaf nodes and then passed up to internal nodes representing the classes

McCarthy s response to the DAG problem is to leave the hierarchy as a DAG and argue that since only around % of the nodes in WordNet have more than one parent the resulting tree cut models are unlikely to differ much from the tree case McCarthy also notes that the majority of cases of multiple inheritance occur low down in the hierarchy r o w p t i 4 4 h y each HMM remains the same but the values of the probabilities vary

To give an example consider how the noun *ro* is generated for the object position of In fact since *ro* has more than one sense in WordNet there are numerous paths through WordNet that generate the noun but let us assume that the noun is generated via the food sense The hypernyms of the food sense of *ro* are as follows ⟨`bread`⟩ ⟨`baked_good`⟩ ⟨`foodstuff`⟩ ⟨`food`⟩ ⟨`substance`⟩ ⟨`object`⟩ ⟨`entity`⟩ First a child of the root of the hierarchy is chosen in this case the ⟨`entity`⟩ node according to the transition probabilities associated with the root Then the concept ⟨`object`⟩ is chosen according to the transition probabilities associated with ⟨`entity`⟩

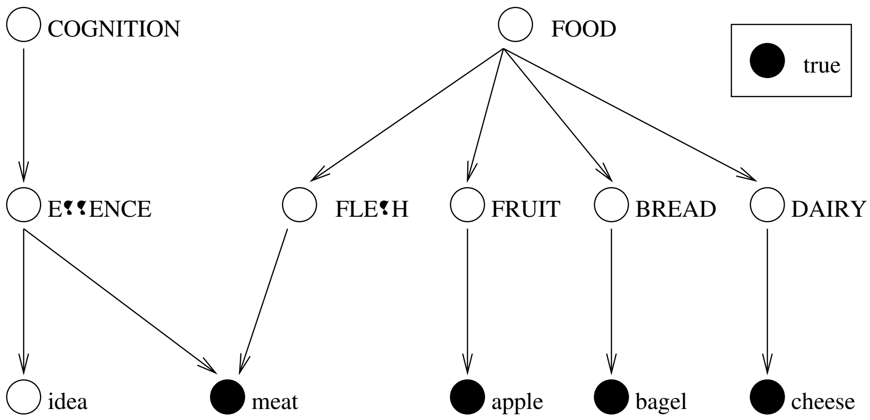


Figure 2 2 Example Bayesian network

variable which can be in one of two states *ru* or *s* A synset node has the value *ru* if the concept represented by the synset is selected for by the verb and a word node has the value *ru* if the word can appear as an argument of the verb

Each variable A with parents $B\ ,\ldots,B_n$ has associated with it a *on on pro* CPT which stores the probabilities $p(A|B\ ,\ldots,B_n)$ Ciaramita and Johnson call these probabil ities the *pr rs* and they are de ned according to the following principles First it is *un* that a verb selects for a concept *pr r* indo word ned

The use of distributional similarity is an important alternative to using a man made hierarchy for generalisation The relevant literature is large and we will only describe some representative approaches Chapter 4 of Manning and chutze also gives an overview of this area After describing a number of approaches we will consider the advantages and disadvantages of using distributional similarity compared with using a man made hierarchy for generalisation

The philosophy underlying distributional approaches is that the probability of a rare event can be estimated by considering similar events that have occurred in the data An example given by Lee and Pereira is that it is possible to infer that the bigram after ACL is plausible even if it does not occur in the data if after ACL does occur in the data This assumes that ACL and ACL have similar cooccurrence distributions or in other words that ACL and ACL tend to occur in the same contexts

imilar events are often organised into clusters according to some probabilistic measure of similarity However as Lee and Pereira point out distributional approaches do not have to explicitly create clusters Dagan Lee and Pereira estimate cooccurrence probabilities by taking the nearest cooccurrences to the target cooccurrence and averaging their probabilities The cooccurrence can be between the head words in a syntactic construction or between words in an n gram for example Lee and Pereira call this approach *n r s n ors r h*

Following Dagan et al let $(\; , \; ')$ be a measure of the similarity between words and $'$ and let $\mathbf{S}(\;)$ be the set of words most similar to ; then $p(\; _2 | \;)$ can be estimated as follows

$$p(\; _2 | \;) = \frac{\sum_{' \in \mathbf{S}(\;)} (\; , \; ') p(\; _2 | \; ')}{\sum_{' \in \mathbf{S}(\;)} (\; , \; ')} \qquad 2 \; 4$$

The numerator is the probability of $_2$ given a nearest neighbour of weighted by a function of the similarity between and the neighbour summed over all the nearest neighbours; and the denominator is a normalising constant

There are a number of similarity measures so rather than attempt to describe them all we use one measure based on the Kullback Leibler KL divergence as an example To measure the similarity between two words and $'$ the KL divergence can be applied as follows

$$\mathrm{D}(\; \| \; ') = \sum_{2} p(\; _2 | \;) \log \frac{p(\; _2 | \;)}{p(\; _2 | \; ')} \qquad 2$$

$\mathrm{D}(\; \| \; ')$

C us r h

Pereira Tishby and Lee acquire clusters of nouns for the direct object position of verbs The clustering is soft in that each word belongs to a cluster according to a cluster membership probability and it is also hierarchical in that the clustering algorithm works in a top down iterative fashion splitting existing clusters at each iteration The decision to keep two nouns in the same cluster is based on the difference between their conditional verb distributions $p_n(\ )$ which is measured using the KL divergence

In contrast Brown Della Pietra de ouza Lai and Mercer 2 adopt a bottom up iterative approach in which initially the clusters are the individual words themselves and the decision to merge two classes is based on the minimal loss of mutual information The clustering is hard in that a noun either belongs to a cluster or it does not and there is no notion of degrees of membership The clustering model was used to try and improve a language model although no improvements in perplexity were gained by using a cluster b *no* *qu s on*

u u n or n

The mutual information between two words x and in some cooccurrence relation is de ned as follows

$$(x,\ ) = \log_2 \frac{p(x,\ )}{p(x)p(\ )} \qquad 2$$

The mutual information described here is often referred to as *po n u u n or n* to distinguish it from the notion used in information theory Pointwise mutual information is derived from the information theoretic notion but the information theoretic version is de ned as an av erage over random variables Also the pointwise version has less of a theoretical basis; Jelinek warns that interpreting $(x,\ )$ as the mutual information between x and gives only an intuitive interpretation p 4

Pointwise mutual information compares the joint probability of observing x and together p

(, 2)	(¬ , 2)
(,¬ 2)	(¬ ,¬ 2)

Table 2 Contingency table for the bigram 2

(, 2) is the number of times 2 follows in the data and (¬ , 2) is the number of times 2 follows a word other than in the data The other frequencies in the table are de ned anal ogously The null hypothesis corresponding to the table is that and 2 appear independently of each other and a statistic such as chi squared can be used to determine how likely the null hy pothesis is to be true If the chi squared statistic has a high value then this gives strong evidence that the null hypothesis is false and that and 2 are highly associated Thus bigrams with high chi squared scores should correspond to highly associated pairs of words or collocations

The chi squared statistic that is usually encountered in text books is the *rson* chi squared statistic However the problem with this statistic as Dunning demonstrates is that it can over estimate the signi cance of rare events This means that the bigrams producing the highest scores are often based on very low counts which makes the test unreliable Most of the top ranked bigrams in Dunning s experiments occurred only once in the data and among the highest ranked bigrams were cases like *pr r n hs n* and *s nn r s* which are hardly highly associated pairs of words As a remedy to this problem Dunning considers the log likelihood ratio statistic denoted G^2 which does not over estimate the signi cance of rare events in the same way The top ranking bigrams produced according to this statistic were much more intuitive

Dunning s analysis of his results is based on the following claim that the sampling distribution of G^2 approaches chi squared quicker than the sampling distribution of X^2 However this part of Dunning s analysis is debatable since Agresti makes exactly 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 p 4

Given Aresti s comments a more likely explanation lies in the conservative nature of G^2 which means that X^2 is more likely to return a signi cant result for a table based on small counts This would explain Dunning s results in which pairs of words occurring infrequently in the corpus obtain high scores according to X^2 but not G^2 These issues will be discussed further in Chapter where a chi squared test is used as part of a procedure for selecting a suitable level of abstraction in WordNet

Pedersen suggests using Fisher s exact test Agresti for bigram discovery rather than a chi squared statistic The advantage of Fisher s exact test is that it can be applied to any contingency table regardless of the size of the counts and the result will be reliable However the test is computationally expensive since it involves computing every contingency table that could have led to the marginal totals observed in the sampled table The marginal totals are not shown in Table 2 but are simply the totals obtained by summing the scores in each row and column In addition the results obtained by Pedersen for the exact test did not differ greatly from those obtained for the log likelihood statistic and so it is not clear that the bene ts of using the test outweigh the additional computational burden

C C

Many of the smoothing techniques used in corpus based NLP were developed for language mod elling and so to demonstrate some of the most widely used techniques we consider the problem of estimating an n gram model More speci cally the problem is to estimate the probability of a word conditional on the previous $n -$ words $p(\, _{n+} \ldots)$ A maximum likelihood

As an example consider using 2 22 to estimate $p(\langle\texttt{fox}\rangle|run,\text{subj})$ and $p(\langle\texttt{carpet}\rangle|run,\text{subj})$ assuming that neither $\langle\texttt{fox}\rangle$ nor $\langle\texttt{carpet}\rangle$ appear with run in the data Unlike additive smoothing the two unseen senses are unlikely to receive the same estimate since the estimates based on less context are unlikely to be the same for the two senses However $\langle\texttt{fox}\rangle$ will not necessarily receive a higher estimate than $\langle\texttt{carpet}\rangle$; the problem is that the estimates based on less context ignore the verb In contrast the estimation method presented in Chapter is able to make use of the verb by determining whether semantically similar senses to $\langle\texttt{fox}\rangle$ and $\langle\texttt{carpet}\rangle$ appear as subjects of run

Goo ur

Another widely used technique is the Good Turing method Good which states that an n gram that has occurred r times in the data should have an adjusted frequency r^* where

$$r^* = (r+\ )\frac{E(\ \ _{r+}\ )}{E(\ \ _r)} \quad (r \geq\ ) \qquad 2\ 2$$

$E(\ \ _r)$ is the expected number of n grams that occur r times in the data Relative frequencies based on the r^* values can be used to estimate the probabilities Note that 2 2 only applies to values of r greater than zero; a further result of Good is that the total probability mass assigned to unseen objects is $E(\ \ )/$ where is the total number of n grams in the data

In practice the actual number of n grams that occur r times in the data n_r can be used to approximate the expected values if the actual values are suitably smoothed themselves To see

This section describes previous work on structural disambiguation which is a problem considered later in the thesis The section describes work on PP attachment and then work that has considered the more general problem of parse selection Not all previous approaches are considered since the literature in both cases is very large and we describe only those approaches that are most relevant to the work in this thesis

The type of structural ambiguity that has been most covered in the literature is PP attachment am biguity This is a pervasive form of ambiguity and a potentially damaging one in that increasing the number of PPs in a sentence can lead to a combinatorial explosion in the number of possible analyses Church and Patil 2 A number of early studies in the psycholinguistics domain sug gested possible strategies for resolving attachment ambiguities Two of the most cited studies are those of Kimball who suggested that a constituent tends to attach to another constituent immediately to its right right association and Frazier who suggested that there is a pref erence for attachments that lead to the parse tree with the fewest nodes minimal attachment However later work Whittemore Ferrara and Brunner ; Taraban and McClelland demonstrated that lexical information is a better predictor of attachments and most of the recent corpus based approaches to structural disambiguation including PP attachment have been based on lexical associations

The PP problem that is usually addressed only considers sequences of the following form *r o o r pr pos on o o pr pos on* Moreover only the heads of the noun phrases are usually considered The problem can then be characterised as as taking a four tuple $(,n ,pr,n_2)$ and deciding whether the PP attaches to or n as in the much used example *s n s op* Note that this is an easier problem than the most general form of PP attachment since only two possible attachment sites are being considered In the general case there may be more than two sites Consider this example from Hindle and Rooth

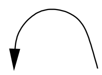
2 24 NBC was so afraid of hostile advocacy groups and unnerving advertisers that it shot its dramatization of the landmark court case that legalised abo

$$p(A| \ ,n \ ,pr,n_2) = \quad \text{if } A \text{ is noun attach} \quad \text{if } A \text{ is verb attach}$$

An interesting result of the paper is that the optimum value for was found to be zero at all stages This means that even if a context occurs only once in the training data it is better to use an estimate based on that context rather than back off to another level We present a related result in Chapter regarding the use of low count events in the training data We nd that simply compares probabilities corresponding to the possible attachment sites An advantage of our approach is that these probabilities can be easily integrated into a model for parse selection

The problem of parse selection is to select the correct parse for a sentence from a number of al ternatives As Collins ; p points out this can be an astonishingly severe problem in broad domains such as the Wall Street Journal WSJ Collins cites a number of factors that are responsible for the severity of the problem the need for a large grammar to obtain broad coverage; long sentences being typical in a broad domain; and many common sources of syntactic ambigu ity such as PP attachment leading to exponential growth in the number of analyses relative to sentence length There are many examples in the literature of ordinary looking sentences having hundreds sometimes thousands of different analyses according to some grammar The parser of

C p r _ r ous or



r ppro s o s p rs

Briscoe and Carroll de ne a probability model based on the moves of an LR parser see also Briscoe and Carroll Carroll and Briscoe Carroll Minnen and Briscoe The grammar underlying the parser is a hand written phrase structure grammar The probability model is structural and does not account for the probabilities of lexical dependencies However more context is taken into account than a PCFG since the history that is considered at each parsing de cision is conditional on the LR state which can encode information in addition to the non terminal being expanded A dependency based evaluation in Carroll Minnen and Briscoe shows that the latest version of the parsing system can identify some grammatical relations such as subject and direct object with high accuracy but is less successful with other relations such as the sec ond object in a ditransitive construction and indirect object The accurate identi cation of some relations such as those corresponding to PP attachment is likely to require a more lexicalised probability model

A current version of the Briscoe and Carroll parser is used throughout this thesis The parser is highly robust and has been used to provide large amounts of training data for the experiments reported in Chapters and It was also used for the parse selection experiments in Chapter in order to provide the possible parses for a set of test sentences A feature of the latest version is that the output is in the form of head dependency relations which were used to create a dependency structure for each possible parse In addition the performance of the parser provided a useful benchmark against which to measure the performance of the dependency model

Hektoen de nes a probability model over logical forms rather than syntactic structures arguing that semantic relations are the key to accurate parse selection A hand written grammar was developed especially for this work so that the requisite logical forms could be derived A further novel aspect of the approach is that Bayesian estimation is used to estimate the parameters Hektoen did attempt a direct comparison with SPATTER and Collins conditional model although the use of a hand written grammar meant that only a subset of sentences from the Penn Treebank could be parsed Also Hektoen argues that the Parseval measures are not very suitable for his system since they measure the ability of the oe nkgue m 2 TJ 2 sinccnld Cm

The problem addressed in this chapter is how to estimate $p(c| \ , r)$ where c is a sense in a semantic hierarchy is a predicate and r is an argument position The term predicate is used loosely here in that the predicate does not have to be a semantic object bu

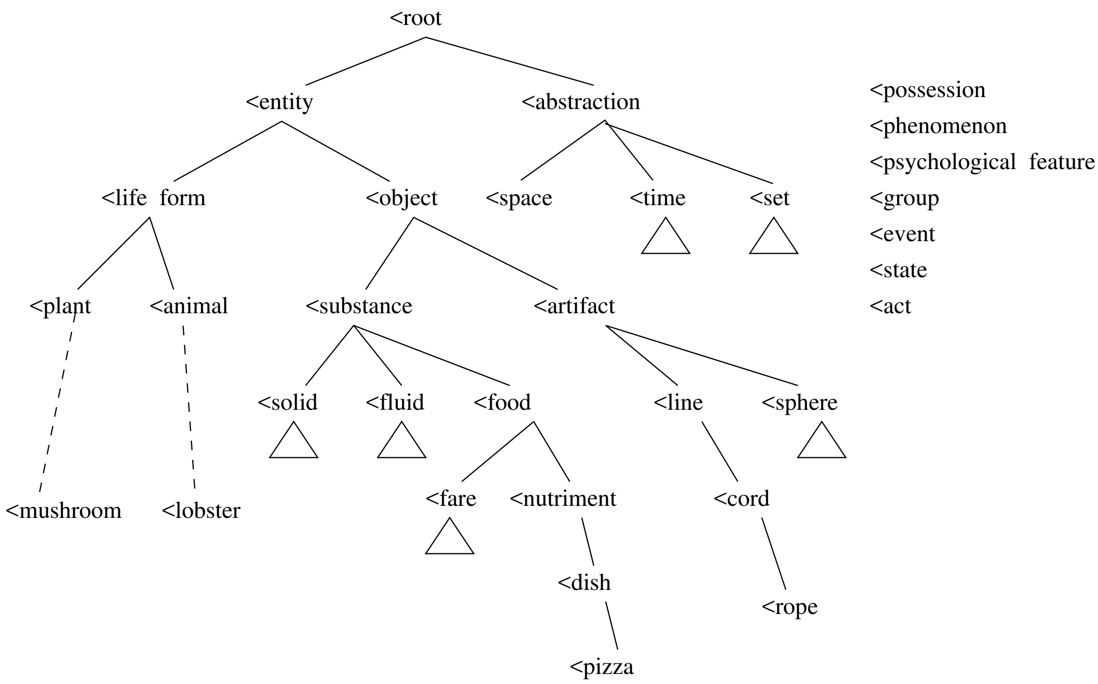


Figure Part of the WordNet hierarchy

concept c and $\mathsf{cn}(n) = \{ c \mid n \in \mathsf{syn}(c) \}$ to denote the set of concepts that can be denoted by the noun n

The hierarchy has the structure of a directed acyclic graph although only around one percent of the nodes in the graph have more than one parent where the edges of the graph constitute what we call the $\mathsf{direct-isa}$ relation Let $\mathsf{isa} \subseteq \mathbf{C} \times \mathbf{C}$ be the transitive re exive closure of $\mathsf{direct-isa}$ then c isa c' implies c is a kind of c' If c isa c' then c is a *p* of c' and c' is a *pon* of c In fact the hierarchy is not a single hierarchy but consists of nine separate sub hierarchies The sub hierarchies are headed by the most general kind of concept and the roots of the sub hierarchies are shown in Figure which shows part of the WordNet hierarchy The seven roots in addition to non verbal predicates such as adjectives as well as verbs

$$
\begin{aligned}
p(\ |\overline{c}, r) &= p(\overline{c}|\ , r)\frac{p(\ |r)}{p(\overline{c}|r)} \\
&= \frac{p(\ |r)}{p(\overline{c}|r)} \sum_{c'' \in \overline{c}} p(c''|\ , r) \\
&= \frac{p(\ |r)}{p(\overline{c}|r)} \sum_{c'' \in \overline{c}} p(\ |c'', r)\frac{p(c''|r)}{p(\ |r)} \\
&= \frac{}{p(\overline{c}|r)} \sum_{c'' \in \overline{c}} \quad p(c''|r) \\
&= \frac{}{p(\overline{c}|r)} \sum_{c'' \in \overline{c}} p(c''|r) \\
&=
\end{aligned}
\qquad 4
$$

Figure 2 Proof of proposition

compare the probabilities $p(\ |\overline{c}_i, r)$ only The proof of proposition is given in Figure and is explained in detail below

The rst line 2 applies Bayes theorem to the probability $p(\ |\overline{c}, r)$ Line rewrites the probability $p(\overline{c}|\ , r)$ as the sum of the probabilities of the sets dominated by the daughters of c $\sum_i p(\overline{c}_i|\ , r)$ plus the probability of c itself $p(c|\ , r)$ This equality holds because the probability of a set of concepts $p(\overline{c}|\ , r)$ has been de ned in as the sum of the probabilities of the concepts in the set However note that the equality only holds in the tree case and this is where the proofs in Figures 2 and differ For a DAG the probability of a set of concepts dominated by c cannot be obtained by summing the probabilities of the sets dominated by the daughters of c plus the probability of c itself The reason is that in the sum $\sum_i p(\overline{c}_i|\ , r)$ the probabilities of

$$
\begin{aligned}
p(\,|\overline{t},r) &= p(\overline{t}|\,,r)\frac{p(\,|r)}{p(\overline{t}|r)} && 2\\
&= \frac{p(\,|r)}{p(\overline{t}|r)}\left(\sum p(\overline{t}\,|\,,r)+p(t|\,,r)\right)\\
&= \frac{p(\,|r)}{p(\overline{t}|r)}\left(\sum p(\,|\overline{t},r)\frac{p(\overline{t}|r)}{p(\,|r)}+p(\,|t,r)\frac{p(t|r)}{p(\,|r)}\right) && 4\\
&= \frac{}{p(\overline{t}|r)}\left(\sum \; p(\overline{t}|r)+\; p(t|r)\right)\\
&= \frac{}{p(\overline{t}|r)}\left(\sum p(\overline{t}|r)+p(t|r)\right)\\
&=
\end{aligned}
$$

Figure Proof of proposition

su 4 o n 2 I IM true W H BPC ID f 2 2 2 2 Tm c 4 TJ 4 2 T

37e9 10.903.68 0 0 0 1 Tf 1 0 0 1 111 08 0 1 312.96 670 [(p..96 60 1617) 1f0 0 1 283.68 0 0e24.966 58986(3)5(,)r28 Tm

	$(, run, \text{subj})$	$(, \text{subj}) - (, run, \text{subj})$	$(, \text{subj}) = \sum_{\in \mathbf{V}} (, , \text{subj})$
⟨bitch⟩		2 2	2
⟨dog⟩	2	2 4 22	2
⟨wolf⟩		4	
⟨jackal⟩		2	2
⟨wild_dog⟩			
⟨hyena⟩	2		
⟨fox⟩	2	2	2
	4		4

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

senses but the data consist of nouns For now a simple approach is taken which is to estimate $(, , r)$ by distributing the count for each noun n in $\mathsf{syn}( )$

C top(, ,r)
top ← c
sig_result ← false
CC parent gives lowest G^2 value G^2
not sig_result top ≠ ⟨root⟩
G^2 ← ∞
parents of top
calculate G^2 for sets dominated by children of parent
$G^2 < G^2$
G^2 ← G^2
parent ← parent

chi squared test for parent is signi cant
sig_result ← true
move up to next node top ← parent

return top

Figure 4 An algorithm for determining top(, ,r)

top(, ,r)

Figure gives an example of the procedure at work Here top(⟨soup⟩,s,obj) is being determined The example is based on data from a subset of the BNC which had cases of an argument in the object position of s The G^2 statistic is used together with an α value of . Initially top is set to ⟨soup⟩ and the probabilities corresponding to the children of ⟨dish⟩ are compared $p(s|\overline{\langle\mathrm{soup}\rangle},\mathrm{obj})$ $p(s|\overline{\langle\mathrm{lasagne}\rangle},\mathrm{obj})$ $p(s|\overline{\langle\mathrm{haggis}\rangle},\mathrm{obj})$ and so on for the rest of the children The chi squared test results in a G^2 value of 4. compared to a critical value of . ince G^2 is less than the critical value the procedure moves up to the next node This continues until a signi cant result is obtained which rst occurs at ⟨substance⟩ when comparing the children of ⟨object⟩ Thus ⟨substance⟩ is the chosen level of generalisation

Before giving some example levels of generalisation it is worth making some comparisons with the other WordNet approaches First note that we have not made a uniform distribution as sumption as Li and Abe do equation 2 Furthermore the problem described in ection 2 stemming from the fact that Li and Abe compare frequencies in order to generalise does not arise This problem is avoided because we compare probabilities conditioned on sets of concepts rather than the frequencies of senses And nally the generalisation procedure is able to return a suitable class for arguments that are negatively associated with some predicate ection 2 explained how such arguments cause a problem for Resnik s approach To see why consider applying the generalisation procedure to ⟨location⟩ in the object position of ; the procedure is unlikely to get as high as ⟨entity⟩ as we argued Resnik s approach is likely to do since the probabilities corresponding to the daughters of ⟨

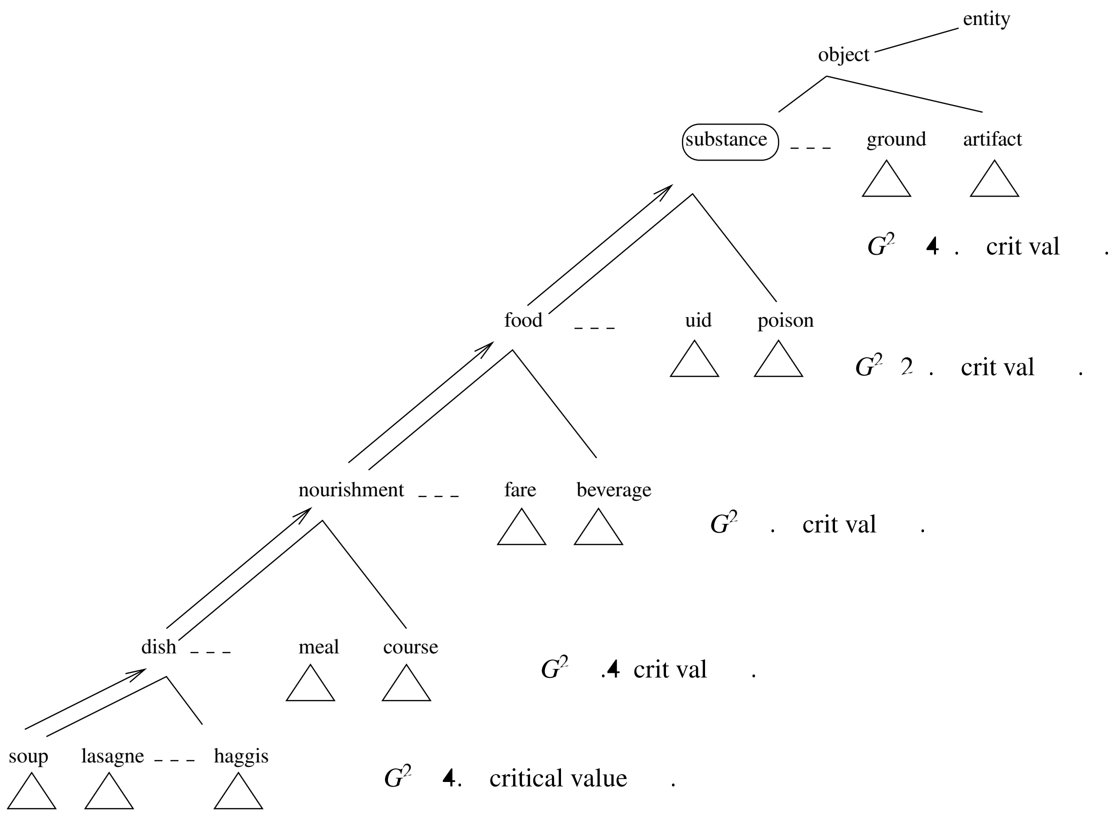


Figure An example generalisation determining top(⟨soup⟩, s, obj)

1 . _C

In this section we show how the level of generalisation varies with the value for α and how

, r) ,r)	α	
⟨coffee⟩, r n ,obj	.	⟨coffee⟩⟨BEVERAGE⟩⟨food⟩...⟨object⟩⟨entity⟩
	.	⟨coffee⟩⟨BEVERAGE⟩⟨food⟩...⟨object⟩⟨entity⟩
r n ,obj) = 4	.	⟨coffee⟩⟨BEVERAGE⟩⟨food⟩...⟨object⟩⟨entity⟩
	.	⟨coffee⟩⟨BEVERAGE⟩⟨food⟩...⟨object⟩⟨entity⟩
⟨hotdog⟩, ,obj	.	⟨hotdog⟩⟨sandwich⟩⟨snack_food⟩⟨DISH⟩...⟨RTfTmT⟩jRTfTmTjRTfaTmTJRTcRTfTfR

,r) ,r)	%	
⟨coffee⟩, r h ,obj		⟨coffee⟩⟨BEVERAGE⟩⟨liquid⟩⟨fluid⟩...⟨object⟩⟨entity⟩
		⟨coffee⟩⟨BEVERAGE⟩⟨liquid⟩⟨fluid⟩...⟨object⟩⟨entity⟩
r h ,obj) = 4		⟨coffee⟩⟨beverage⟩⟨liquid⟩⟨FLUID⟩...⟨object⟩⟨entity⟩
		⟨coffee⟩⟨beverage⟩⟨liquid⟩⟨fluid⟩...⟨object⟩⟨entity⟩⟨ROOT⟩
⟨hotdog⟩, ,obj		⟨hotdog⟩...⟨DISH⟩⟨nourishment⟩⟨food⟩...⟨entity⟩
		⟨hotdog⟩...⟨DISH⟩⟨nourishment⟩⟨food⟩...⟨entity⟩
,obj) = ,		⟨hotdog⟩...⟨dish⟩⟨NOURISHMENT⟩⟨food⟩...⟨entity⟩
		⟨hotdog⟩...⟨dish⟩⟨nourishment⟩⟨food⟩...⟨entity⟩⟨ROOT⟩
⟨Socrates⟩, s,obj		⟨Socrates⟩...⟨life_form⟩⟨CAUSAL_AGENT⟩⟨entity⟩
		⟨Socrates⟩...⟨life_form⟩⟨CAUSAL_AGENT⟩⟨entity⟩
s,obj) = 4		⟨Socrates⟩...⟨life_form⟩⟨causal_agent⟩⟨ENTITY⟩
		⟨Socrates⟩...⟨life_form⟩⟨causal_agent⟩⟨entity⟩⟨ROOT⟩
⟨dream⟩,r r,obj		⟨dream⟩...⟨preoccupation⟩⟨cognitive_state⟩⟨STATE⟩
		⟨dream⟩...⟨preoccupation⟩⟨cognitive_state⟩⟨STATE⟩
r r,obj) = , 2		⟨dream⟩...⟨preoccupation⟩⟨cognitive_state⟩⟨state⟩⟨ROOT⟩
		⟨dream⟩...⟨preoccupation⟩⟨cognitive_state⟩⟨state⟩⟨ROOT⟩
⟨man⟩,s ,obj		⟨man⟩...⟨mammal⟩...⟨animal⟩⟨LIFE_FORM⟩⟨entity⟩
		⟨man⟩...⟨mammal⟩...⟨animal⟩⟨LIFE_FORM⟩⟨entity⟩
s ,obj) = ,		⟨man⟩...⟨mammal⟩...⟨animal⟩⟨LIFE_FORM⟩⟨entity⟩
		⟨man⟩...⟨mammal⟩...⟨animal⟩⟨life_form⟩⟨entity⟩⟨ROOT⟩
⟨belief⟩, n on,obj		⟨belief⟩...⟨cognition⟩⟨

α	%	%	%	%
.	.	.	.	.
.	2.	.	4.	.
.	2.	2.	4.	.4
.	.2	.	2.	.

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

α	G^2	X^2
.	.	.
.	2.	2.
.	2.	.
.	.2	.2

G^2 statistic The advantage of this test is that it can be applied to any contingency table irrespective of the size of the counts The main disadvantage 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 with low counts tends to produce an insigni cant result and the null hypothesis is not rejected This is especially true for the more conservative G^2 statistic The consequences of this for the generalisation pro cedure are that low count tables tend to result in the procedure moving up to the next node in the hierarchy This behaviour is clearly demonstrated in Tables 4 and 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 next issue to consider is which statistic to use Dunning argues that G^2 is more suitable for corpus based linguistics than X^2 and Chapter 2 described Dunning s experiment com paring the use of X^2 and G^2 to identify highly associated bigrams Dunning s claim is that for small samples the sampling distribution of G^2 is a better approximation to the chi squared dis tribution than the sampling distribution of X^2 However in Chapter 2 we presented a quotation from Agresti which contradicts this claim A more likely explanation lies in the conservative nature of G^2 which means that X^2 is more likely to return a signi cant result for a table based on small counts This would explain Dunning s bigram results in which pairs of words occurring infrequently in the corpus obtain high scores according to X^2 but not G^2

Note that for some applications it may make little difference to the performance whether G^2 or X^2 is used The results for a PP attachment task described in Chapter are very similar for both statistics In fact the use of X^2 may even lead to better results for some applications The results of a pseudo disambiguation task also described in Chapter plenty of counts; and since the point of this work is to overcome the sparse data problem the second consideration should override the rst The chi squared test has this overriding effect built in automatically particularly when using the conservative G^2 statistic since it measures the

This may appear to be a crude solution to the problem of ambiguous data but in practice it works surprisingly well The reason is that counts for sets of concepts tend to accumulate in the right places To see why consider this example adapted from Resnik Resnik notes that a similar point is made by Yarowsky 2 Consider estimating probabilities for the object position of the verb *rh* and suppose that *rh h* and *rh r* occur as part of the data The word *r* is a member of seven senses in WordNet and *h* is a member of two senses Thus for these data items splitting the count equally leads to each sense of *r* receiving . 4 counts and each sense of *h* . counts But note that with regard to *s s* of concepts only those sets containing senses of both *h* and *r* such as $\langle$`beverage`$\rangle$ will accumulate counts The counts for the incorrect senses will be randomly dispersed throughout the hierarchy as noise and areas where counts would be expected to accumulate such as under $\langle$`beverage`$\rangle$ in this example will receive the majority of the overall count As will be shown later this accumulation effect means that performance in applications can be good even if this simple estimation technique is used

However there is an obvious problem with this approach although counts for sets tend to accumulate in the right places counts can be greatly underestimated In the previous example $(\overline{\langle\texttt{beverage}\rangle}$, *rh* ,obj$)$ is incremented by only . 4 counts from the two data instances rather than the correct value of 2 In addition as Resnik himself notes the accumulation process has less effect on sets of concepts low down in the hierarchy since here the counts have had less chance to accumulate The example Resnik gives is for *o nos* In this case counts would be expected to be higher for the set dominated by the bodily sense of *nos* rather than the aircraft sense However since both senses are low down in the hierarchy splitting counts equally is likely to lead to a similar count for each set For the same reason counts for individual concepts as opposed to sets of concepts are likely to be inaccurate

In response to this we note that the accumulation of counts leads to an obvious strategy use the fact that correct senses are likely to be members of sets where counts have accumulated as a way of re distributing the count Continuing with the *rh h* example *h* has a beverage sense and a colour sense in WordNet If the above strategy is used equal counts will be given to each sense on the rst iteration but on subsequent iterations more of the count will be given to the beverage sense This is because counts would accumulate under $\langle$`beverage`$\rangle$ for the object position of *rh* and not under $\langle$`colour`$\rangle$

One issue to consider is how to determine a representative set for a concept We have been assuming that $\overline{\langle\texttt{beverage}\rangle}$ and $\overline{\langle\texttt{colour}\rangle}$ are suitable for the two senses of *h* but a procedure is needed which determines this automatically The procedure needs to nd a hypernym for each alternative sense such that the hypernym is high enough for counts to have accumulated in the set dominated by the hypernym; however it should not be so high that the alternative senses cannot be distinguished An example of a hypernym that is too high is $\langle$`root`$\rangle$ the notional root of the hierarchy since if $\langle$`root`$\rangle$ were chosen for both senses of wine there would be no way to distinguish between the senses Another reason not to go too high is that the sets need to be in some sense representative of the senses uppose *p* occurs in the data and the food sense of *p* and the electronic sense need to be distinguished It would not be appropriate to represent the electronic sense using $\langle$`entity`$\rangle$ since this would not capture the rto 2 t 2 2 h a 2

$$\mathrm{A}\ \ (C, \ ,r) = \frac{p\ \ (C|\ ,r)}{p\ \ (C|r)}$$

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

$$p\ \ (C|r) = \frac{\sum_{\ \in \mathbf{V}}\ \ (C,\ ,r)}{\sum_{\ \in \mathbf{V}}\ \ (\ ,r)}$$

$$\ \ (C,\ ,r) = \sum_{\in C}\ \ (\ ,\ ,r)$$

Figure 4 2 Estimates for calculating A $(C, \ ,r)$ for a set of concepts C; $\mathbf{V}$ is the set of verbs in the data

$\overline{\langle \texttt{entity} \rangle}$ is not homogeneous with respect to the object position of rh some entities are drunk some are not In contrast the set $\overline{\langle \texttt{abstraction} \rangle}$ is fairly homogeneous in that on the whole kinds of abstraction are rarely drunk

The set $\overline{\langle \texttt{beverage} \rangle}$ is also homogeneous which is a suitable representative for the beverage sense Note that the two sets $\overline{\langle \texttt{abstraction} \rangle}$ and $\overline{\langle \texttt{beverage} \rangle}$ are also maximally homogeneous in that the sets dominated by the parents of $\langle \texttt{beverage} \rangle$ and $\langle \texttt{abstraction} \rangle$ $\overline{\langle \texttt{liquid} \rangle}$ and $\overline{\langle \texttt{root} \rangle}$ respectively are not themselves homogeneous This motivates the idea that we should be looking for maximally homogeneous sets maximal because we want to allow counts to accumulate and noise to be dispersed The problem with using $\overline{\langle \texttt{colour} \rangle}$ as a representative of $\langle \texttt{wine} \rangle$ is that $\langle \texttt{colour} \rangle$ is not high enough for this dispersal to have occurred

One way to recognise that $\overline{\langle \texttt{liquid} \rangle}$ is not homogeneous is to note that the sets dominated by the daughters of $\langle \texttt{liquid} \rangle$ are associated to differing degrees with rh ome liquids are drunk such as beverages liquor and water but some are not such as ammonia antifreeze and sheep the verb Thus it appears that the procedure can be applied directly to the problem of determining [, r]

However there are some differences between the problems being addressed in this and the previous chapter In the previous chapter the problem was to nd a generalisation level that would lead to a reasonable probability estimate In this chapter the problem is to nd a level where counts have accumulated and the noise dispersed suf ciently A solution to both problems lies in nding homogeneous sets; the difference lies in the r of homogeneity that is likely to be optimal in each case For the probability estimation problem it may be that the difference in association norms needs to be relatively small for a class based probability estimate to be a useful estimate Results presented in Chapter suggest that for some disambiguation tasks this is indeed the case Another way to think of this is that for some tasks the optimal level of generalisation is quite low in the hierarchy on the whole In contrast the re estimation problem is likely to favour a level of generalisation that is quite high on the whole since it is here that counts have accumulated and noise dispersed

Despite these differences the procedure can be adapted to both problems The degree of homogeneity required can be controlled by the parameter α the level of signi cance of the chi squared test The value of α controls the overall level of generalisation a high value for α results in a low level of generalisation on the whole and a low value for α results in a high level of generalisation Results from the previous chapter clearly demonstrate this One way to set a value for α would be to estimate counts using a range of α values and use a held out test set to choose those counts that give the best performance on the task in hand

Another useful feature of the procedure within the context of the re estimation problem is that it employs a signi cance test to nd homogeneous sets This implies that the procedure automatically nds areas where counts have accumulated since it is only here that there will be enough data to return a signi cant result for the chi squared test This point is especially true when the more conservative G^2 statistic is used and a low value for α

As a nal comment a point of clari cation is needed The previous chapter showed that the chosen level of generalisation is dependent on the size of the data sample as well as on the value of α Thus the notion of homogeneity being used here is not an absolute notion but a relative one relative to the sample If the procedure determines a maximally homogeneous set that does not accord with intuition this should not be automatically considered a failure A comment in Clark and Weir states that $\overline{\langle\texttt{food}\rangle}$ is heterogeneous with respect to the object position of

$$p(\overline{\langle\texttt{food}\rangle}| \quad ,\text{obj}) = \frac{(\overline{\langle\texttt{food}\rangle}, \quad ,\text{obj})}{(\quad ,\text{obj})}$$
$$= \quad /2, \; 4$$
$$= \; . \; 4$$

$$p(\overline{\langle\texttt{food}\rangle}|\text{obj}) = \frac{(\overline{\langle\texttt{food}\rangle},\text{obj})}{(\text{obj})}$$
$$= \quad , \quad / \; , \quad ,$$
$$= \; . \quad 2$$

$$\mathrm{A}(\overline{\langle\texttt{food}\rangle}, \quad ,\text{obj}) = \; . \; 4 \; / \; . \quad 2$$
$$=$$

Figure 4 4 Calculation of $\mathrm{A}(\overline{\langle\texttt{food}\rangle}, \quad ,\text{obj})$

high value for $(\overline{\langle\texttt{entity}\rangle}, r)$ and so $p(\text{cajole}|\overline{\langle\texttt{entity}\rangle},\text{obj})$ is not over estimated

The conclusion is that if the association norm is to be applied appropriately it should be applied to frequent verbs or to sets for which (C,r) is reasonably high; however since the re estimation procedure relies on using sets where plenty of counts have accumulated this should not be a problem

1

There are two evaluations in this section[4] The rst shows how the estimated counts change

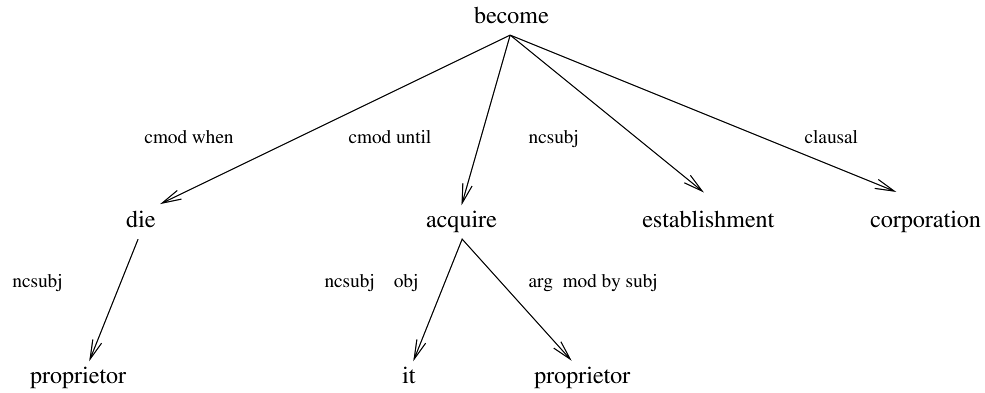


Figure Example dependency structure for the sentence *n propr or s s s n s ou o orpor on un quir no r propr or*

o un quir the establishment should *o* a corporation *un* it is *quir* by another proprietor Here *o* is the head in both cases and and *quir* are dependents The prepositions *n* and *un* introduce the dependents

- ncsubj denotes a non clausal subject The ncsubj examples simply encode a head and dependent except that the passive *s quir* is recognized as such by the symbol obj This appears in the triple labelling the edge *quir* and indicates that is an underlying object of *quir*
- arg_mod

Generate the non dependent heads Θ
head in Θ
Generate a bag of grammatical relations
relation in bag
Generate a transformation
Generate a dependent and type introducing the dependent

the leaves of the generated structure are all null dependents
non null leaf dependent
Generate a bag of grammatical relations
relation in bag
Generate a transformation
Generate a dependent and type introducing the dependent

Figure 2 equence of decisions generating a dependency structure

The dependency structure with the highest probability is chosen as the correct structure together with the corresponding parse if necessary The conditioning context $\ldots$ is known as the history and is equivalent to the structure built up to that point In order that the model have a manageable number of parameters a function Φ

$p(\ , \ | \ , r)$ where is a nominal dependent

The probabilities corresponding to the above examples are

- $p(\ \ p \ \ r on | p \ \ , \mathsf{iobj})$
- $p(\ \ r \ \ or \ | r \ \ , \mathsf{ncmod})$
- $p(\ on \ on \ h | \ \ h \ , \mathsf{ncmod})$

Again the sense of is chosen which maximises the probability estimate and $p(\ , \ | \ , r)$ is used as a proxy for $p(\ , \ | \ , r)$ where is determined as follows

$$= \arg \max_{\in \mathsf{cn}(\ )} p_s(\ ', \ | \ , r)$$

The class based approach can be used to obtain $p_s(\ ', \ | \ , r)$ by rst applying Bayes theorem and then conditioning on an appropriate set of concepts as before The only difference is that the conditional probability of is now joint with

$$\begin{aligned} p(\ ', \ | \ , r) &= p(\ , \ | \ ', r) \frac{p(\ ' | r)}{p(\ | r)} \\ &\approx p(\ , \ | \overline{ \ '}, r \end{aligned}$$

The set $\overline{c'}$ is obtained by applying the procedure described in Chapter and the probability $p(\;|\overline{c'},r)$ is estimated using relative frequencies If the head does not appear in WordNet an estimate of $p(\;|\overline{\langle\texttt{root}\rangle},r)$ is used unless the head is a pronoun or proper name If the head is a pronoun $\overline{c'}$

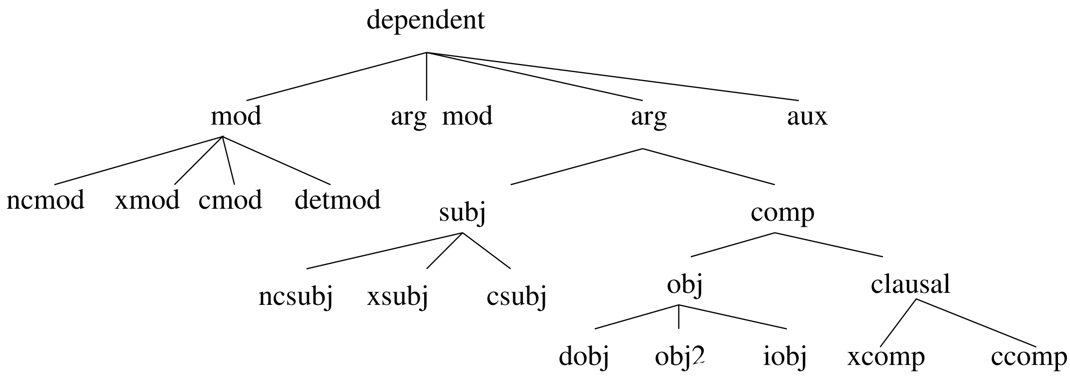


Figure The grammatical relations used in the implementation

1 C C

1 CC

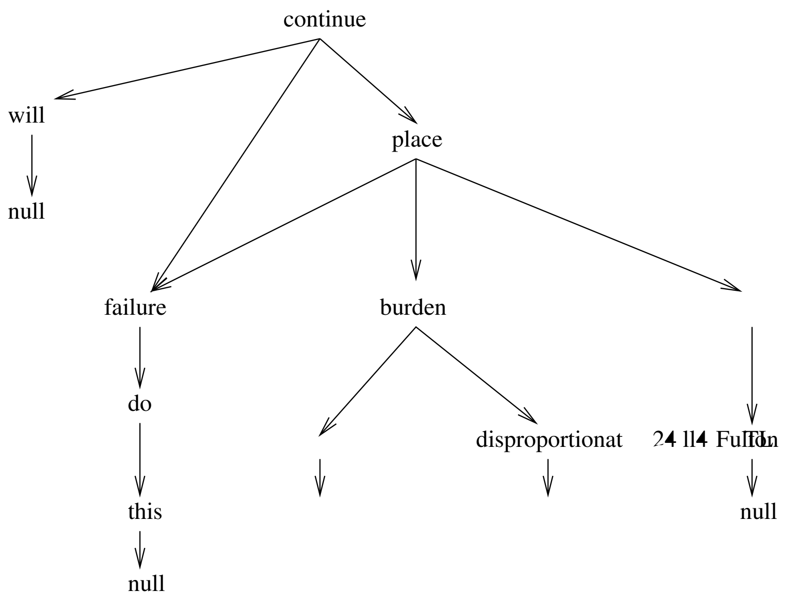
The parser used for the evaluation is a more developed version of that described in Carroll and Briscoe This version is able to produce output in the form of grammatical relations which is the main reason the parser was chosen The parser produces a set of parses for a sentence together with the corresponding sets of grammatical relations Thus we were able to create a dependency structure for each parse and choose the parse with the most probable structure A further advantage in using this parser is that there exists a manually created test suite which uses the same grammatical relation scheme as used by the parser Carroll et al a ; this test suite was used for the evaluation

The relations used by the parser can be arranged in a hierarchy as shown in Figure If the parser is unable to determine the precise nature of the relation and thus cannot return a relation at a leaf node a more generic relation can be returned Each relation is described in detail in Appendix A based on the descriptions given in Carroll et al a and Carroll et al A brief description of each relation is given below

•

```
(|ncsubj| |continue:6_VV0| |failure:1_NN1| _ )
(|clausal| _ |continue:6_VV0| |place:8_VV0|)
(|ncsubj| |place:8_VV0| |failure:1_NN1| _ )
(|dobj| |place:8_VV0| |burden:11_NN1| _ )
(|iobj| |on:12_II| |place:8_VV0| |tax-payer:14_NN2|)
(|dobj| |do: _VD0| |this:4_DD1| _ )
(|xcomp| |to:2_TO| |failure:1_NN1| |do: _VD0|)
(|ncmod| _ |burden:11_NN1| |disproportionate:10_JJ|)
(|ncmod| _ |tax-payer:14_NN2| |Fulton:1 _NP1|)
(|detmod| _ |burden:11_NN1| |a:15_AT1|)
(|aux| _ |continue:6_VV0| |will:16_VM|)
```

Figure Example parser output for the sentence


continue
will
place
null
failure
burden
do
disproportionat
this
null
null

obtained from John Carroll who ran the parser over around million words of the BNC from around , sentences The parser output was in the same form as that given in Figure and the output was processed in the following way the formulaic expressions such as sums of money were found using simple regular expressions

- 4 digit numbers beginning or 2 were replaced with the word *on*

  Numerical expressions were replaced with *n _qu n*

  Monetary expressions not in WordNet were replaced with *su _o _ on*

  Expressions denoting people not in WordNet such as Dr were replaced with *so on*

  Expressions denoting companies not in WordNet such as Ltd were replaced with *o p n*
- Verbs and prepositions were reduced to lower case
- All words were lemmatized

The formulaic expressions were replaced with these particular words because each word has only one sense in WordNet and belongs to a relevant synset

ome parts of the data are much more accurate than others Table in the next section

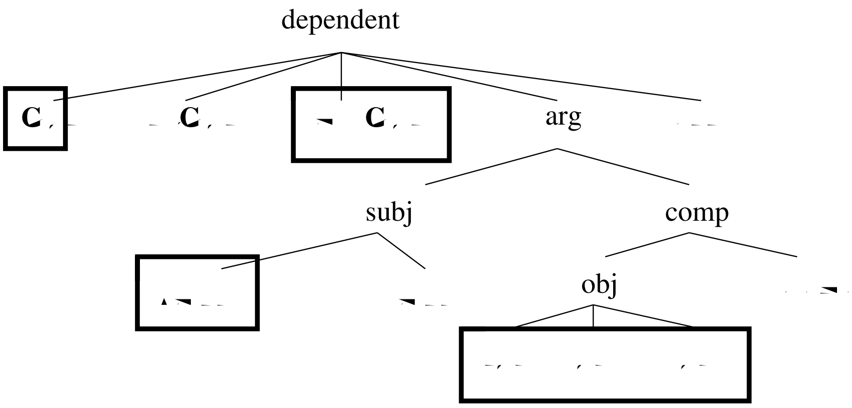


Figure Dependency probabilities by relation that can be estimated using WordNet

is half covered by a box because not all of the mod cases can be estimated using WordNet For the test suite used for the evaluation approximately % of the grammatical relations correspond to parameters that can be estimated using WordNet The parameters corresponding to the remaining relations were estimated using the linear interpolation method

1

The test suite consists of sentences taken from the Susanne corpus covering a number of written genres and manually annotated with grammatical relation information

[redacted]

Relation	occurrences	% occurrences
dependent		.
mod		.
ncmod	24 4	.2
xmod	2	2.
cmod	2	.2
detmod	24	.2
arg_mod	4	.
arg	2	.2
subj	4	.
ncsubj		.
xsubj		.
csubj		.
comp		.
obj		.
dobj	4	.
obj2		.
iobj		2.4
clausal	4 4	.2
xcomp	2	4.
ccomp		.2
aux		.
conj	4	2.

Table Frequency of each type of relation in the test suite

structures The model is likely to prefer incomplete structures with a small number of relations because in these cases less probabilities are multiplied together to get a total probability for the dependency structure

The dependency structures were processed in similar ways to the data in that each word was lemmatized and formulaic expressions were replaced with words in WordNet as described in Section 2 Because there is only a small amount of data in the test set we did not use any of it as held out data and the various parameters were selected by hand The parameters δ and ε described in Section 2 2 were set to , and respectively and the level of signi cance for the chi squared test α was set to . The results appear at t H BPC ID EI Qq 24f signi c

(t)3r0(h)5(e993(i)3.0011(n)-36(t)2.t 10986(s)4(e)-227347.9

Relation	Precision %	Recall %	F score	GRs
dependent	2.	.	.	
mod	.	.2	2.	
ncmod	.	.4	.	2
xmod	.	24.	4.	4
cmod	.	2 .	.	
detmod	.	.2	2.	2
arg_mod				
arg	.	.	2.	2
subj	.	.2	.	2
ncsubj	.	.	.4	
xsubj	.	2 .	.	
csubj				2
comp	.	.	.4	
obj	.	.	.	
dobj	.	2.2	.2	42
obj2	2 .	.	4 .	4
iobj	2.	.4	4 .2	
clausal	.2	.2	.	
xcomp	. 4	2 Tf	2 2	

Tf 22 Tm 42 4 4

The treatment of word sense ambiguity is another area that could be improved Currently a rather cavalier approach is taken which is to select the sense that maximises the relevant probabil ity estimate One promising approach is to try and integrate the word sense disambiguation into the parsing model and perform the two simultaneously as Bikel 2 has attempted to do

A tentative conclusion of this chapter is that the use of lexical sense preferences or selectional preferences alone is unlikely to lead to a highly accurate parse selection system Even the suc cessful statistical parsing models such as those of Collins and Charniak 2 which rely heavily on lexical information also make use of the structural properties of a parse One way to extend this work would be to try and combine the dependency model with the structural model of Briscoe and Carroll

As an evaluation of the class based estimation technique the results are inconclusive since the parse selection problem may not be a good way to isolate the performance of the WordNet estimation techniques In order to have a more focused evaluation the method of estimation is applied to two disambiguation tasks that can be tackled using only parameters relating to lexical sense preferences; moreover the parameters can be estimated using reliable data These tasks are presented in the next chapter

For these examples it is hard to see that there is an ambiguity at all but the attachment problem assumes that any *r np pr p np* sequence results in an ambiguity In it is assumed that *o o p n* could attach to ; in *4 n o n or* could attach to *us*; and in *sp sp u on* could attach to *n s s*

Another reason why the telescope and stick examples are misleading is that they imply the PP attachment problem as we have de ned it is harder than it really is For these two examples either attachment results in a plausible semantic reading and the correct reading depends on the wider context In a commonly cited paper Altmann and teedman argue that the resolution of attachment ambiguities requires a model where the relevant entities are represented and reasoned about This argument led Hindle and Rooth to comment that if this is typical of PP attachment ambiguities then there is little hope of building computational models to solve the problem at least in the near future

Clearly some account of context is required for the resolution of some cases of attachment ambiguity However this may only apply to a small subset of cases The three treebank examples can be resolved without resorting to the wider context; in fact they can be resolved without even considering n_2 ui

The estimates $p_{s}(\, ,pr| \;)$ and $p_{s}(_n \, ,pr|n \;)$ are obtained using the method described in Chapter First Bayes rule is applied and then probabilities are conditioned on a set of concepts where appropriate The formulae are given for $p(\, ,pr|$

4 C p r A so on o p r on o ss s s on n s

α value	% correct G^2	% correct X^2
.	. (4 cases	. (cases
.	. (, 2 cases	

$\max_{\in \mathsf{cn}(n)} p_s($,obj Tj R 2 Tf 2 2 Tm o a 4 x TJ R Tf 4 2 2 Td c T

Generalisation technique	% correct	av gen	sd gen
imilarity class			
$\alpha =$.	.	.	2.
$\alpha =$.	.4	2.	.
$\alpha =$.	.	2.4	.
$\alpha =$.	.	.	.
$\alpha =$.	.	.2	.2
Low class	.	.	.
MDL	.	4.	.
Assoc	.	4.2	2.

Table   Results for the pseudo disambiguation task

it as a noun noun sense pair For example the two instances of *o* in the synsets $\{$ *o* $\}$ and $\{$ *o* , *o* , *o* , *sno* , $C\}$ are treated as separate nouns We use $\mathsf{sep}(n)$ to denote the set of separate instances of n in WordNet

Adopting the MDL approach the disambiguation decision was made as follows p is used to denote an estimate using the MDL approach

$$\max_{n' \in \mathsf{sep}(n)} p(n' | \;,$$

α value	% correct G^2	% correct X^2
.	. (.)	4. (.)
.	.4 (2.)	. (2.)
.	. (2.4)	4. (2.2)
.	. (.)	4. (.)
.	. (.2)	. (.2)

Table Disambiguation results for G^2 and X^2

important feature of these results is that the α values corresponding to the lowest scores lead to a signi cant amount of generalisation This explains why the α

This Chapter considers each of the problems that have been addressed in this thesis outlining the proposed solution for each problem together with the original contribution The ways in which the work could be extended are also considered The discussion is organised by chapter

considered the problem of how to estimate the probability of a noun sense given a predicate and argument position The proposed solution answers two questions one how to use a class from WordNet to estimate the probability of a noun sense thereby overcoming the sparse data problem ; and two how to select a suitable class to represent a sense The second question can be thought of as how to select a suitable level of generalisation in WordNet The proposed generalisation procedure employs a chi squared test and the level of signi cance of the test α is treated as a parameter to be set empirically Results were given showing how the chosen level of generalisation depends on both the sample size and the value of α

The generalisation procedure is arguably the most important contribution of the thesis As Resnik a comments It has been widely noted that the selection of an appropriate level of abstraction is a dif cult problem p We have tried to devise a procedure that has a clearer statistical interpretation than that of Resnik and also one that overcomes some of the shortcomings of Li and Abe s approach such as the uniform distribution assumption 2 An advantage of our approach is that treating α as a parameter gives the procedure a level of exibility since α can be set to produce a level of generalisation that is appropriate for the task in hand

An alternative to using a single class to estimate the probability of a concept which was suggested by Jason Eisner at COLING 2 is to use all the classes dominated by the hypernyms of a concept An estimate would be obtained for each hypernym and the estimates combined in a linear interpolation An approach similar to this is taken by Bikel 2 in the context of statistical parsing

described an unsupervised reestimation algorithm for estimating sense frequencies We rst explained how splitting the count for a noun equally among its senses works better than might be expected at least for the frequencies associated with sets of senses The reason is that counts tend to accumulate in the right places in WordNet namely for sets of senses that are positively associated with the predicate This accumulation effect motivated the reestimation algorithm in which the count for a noun is split equally on the rst iteration but on subsequent iterations more count is given to those noun senses that belong to positively associated sets A feature of the algorithm is that it employs the generalisation procedure described in Chapter and this led to a new interpretation of the procedure as one that nds sets of semantically similar senses or homogeneous sets of senses in the hierarchy The results on a pseudo disambiguation task showed that the reestimation can be bene cial in some cases

The performance of the reestimation algorithm is limited by the fact that highly accurate WSD is unlikely to be achieved using preferences alone Other work that has attempted to use prefer ences for sense disambiguation has achieved little success Resnik ; Carroll and McCarthy 2 Thus one way to further this work would be to see how other knowledge sources could be used to aid the reestimation The surrounding context of a noun is an obvious source of addi tional information There also needs to be more research int the original method of Hindle and Rooth It was discovered that in order to perform well the disambiguation method requires more training data than currently exist in treebanks but that with appropriate amounts of data the method is highly accurate It was also shown that the gen eralisation procedure introduced in Chapter outperforms a simple approach of choosing a xed level in the hierarchy

A further evaluation using a pseudo disambiguation task showed that our class based estima tion method outperforms two alternative approaches based on the work of Resnik a and Li and Abe It was discovered that the alternative methods appeared to be over generalising at least for this task As we have argued a useful feature of our estimation procedure is that the level of signi cance used in the chi squared test α can be used to guard against over or under generalisation But even when the results did vary with α our method was found to outperform the alternatives across the whole range of α values

A further useful result was that the performance on the task was at least as good when using the Pearson chi squared statistic as when using the log likelihood chi squared statistic This result is at odds with the currently accepted wisdom that the log likelihood chi squared statistic is a better statistic for use in corpus based NLP We suggested an explanation for this nding which also explains the results of Dunning who initially argued for the use of the log likelihood statistic

An important question that has yet to be addressed in the literature is whether class based estimation methods perform better when the classes are automatically acquired or when they are part of a man made hierarchy One way to investigate this would be to perform the pseudo disam biguation task but using clustering algorithms to estimate the probabilities Pereira et al and Rooth et al have already used a similar task to evaluate their clustering algorithms; the results depended on the number of clusters induced and ranged between % and % for both approaches compared to the % reported here Unfortunately different test and training data were used in each case and so it is dif cult to draw any conclusions from these 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 would be more conducive to probability estimation

Charniak E 2 A maximum entropy inspired parser In *ro s o s o or A r n C p r o Asso on or Co p on u s* pp 2 eattle WA

Chen and Goodman J An empirical study of smoothing techniques for language modeling In *ro s o Ann o Asso n o*

Collins M and Brooks J Prepositional phrase attachment through a backed off model In *ro h s o r GDA or s op on r r Corpor* pp 2 Cambridge MA

Cover T and Thomas J *E n s o n or on or* Wiley New York

Daelemans W Van Den Bosch A and Zavrel Z Forgetting exceptions is harmful in language learning *h rnh* 4

Dagan I Lee L and Pereira F imilarity based models of word cooccurrence probabilities *h rnh* 4

Dagan I Marcus and Markovitch Contextual word similarity and estimation from sparse data *Co pu r p n n u* 2 2

Drange T *p Crossh s n n nh ss h Bor r Ar o h us s n osop* Mouton The Hague

Dunning T Accurate methods for the statistics of surprise and coincidence *Co pu on h us s* 4

Eisner J a An empirical comparison of probability models for dependency grammar Tech

Eisner r i probabilist c 4 d 24 m 2

di 24 s 4 4 44 o f 2 h r o f u g f 2 I 4 p u t 4t i o4 i t i 2

B r p

Good I J The population frequencies of species and the estimation of population param eters *B r* 4 2 2 4

Goodman J Probabilistic feature grammars In *ro s o n rn on or s op on rs no o s* pp Boston MA

Grishman R Macleod C and Meyers A 4 Comlex syntax Building a computational lexicon In *ro s o n rn on Con r n on Co pu on u s* pp 2 2 2 Kyoto Japan

Harrison P Abney Black E Flickinger D Gdaniec C Grishman R Hindle D Ingria B Marcus M antorini B and trzalkowski T

Joshi A and chabes Y 2 Tree adjoining grammars and lexicalized grammars In Nivat M and Podelski A Eds *D n n o n o s o r s* Elsevier Princeton NJ

Katz J and Fodor J 4 The structure of a semantic theory In Fodor J and Katz J Eds *ru ur o n u* chap pp 4 Prentice Hall Englewood Cliffs NJ

Katz Estimation of probabilities from sparse data for the language model component of a speech recognizer *EEE r ns ons on A ous s p n n ro ss* 4 4

Kilgarriff A Which words are particularly characteristic of a text A survey of statistical approaches In *ro s o A B or s op on n u En r or Do u n An s n o n n* pp 4 ussex UK

Kilgarriff A and Rose T Measures for corpus similarity and homogeneity In *ro s o r Con r n on E p o s ur n u ro ss* pp 4 2 Granada pain

Kimball J even principles of surface structure parsing in natural language

Lin D Automatic retrieval and clustering of similar words In *ro s o* *Ann o Asso or Co pu on u s* pp Montreal Canada

Magerman D 4 *ur n u rs s rn o n on* Ph D thesis University of tanford

Magerman D tatistical decision tree models for parsing In *ro s o r Ann o Asso or Co pu on u s* pp 2 2 Cambridge MA

Prescher D Riezler S and Rooth M 2 Using a probabilistic class based lexicon for

B r p

Ribas F b On learning more appropriate selectional restrictions In ro s o Con r n o Europ n C p r o Asso on or Co pu on u s pp 2 Dublin Ireland

Riezler Prescher D Kuhn J and Johnson M 2 Lexicalized stochastic modeling of constraint based grammars using log linear measures and EM training In ro s o Annu o Asso on or Co pu on u s pp 4 4 Hong Kong

Rooth M Riezler Prescher D Carroll G and Beil F Inducing a semantically annotated lexicon via EM based clustering In ro s o Annu o Asso on or Co pu on u s pp 4 University of Maryland MD

ampson G En or Co pu r Oxford University Press Oxford UK

teedman M 2 n ro ss The MIT Press Cambridge MA

tetina J and Nagao M Corpus based PP attachmen

Some of the descriptions given here are taken directly from Carroll et al a and the same notation is used Many of the examples also come directly from that paper

The relation between a head and a modi er; is used to indicate the word introducing the dependent where appropriate Examples include the following

mod ag red	a red ag
mod with walk John	walk with John
mod while walk talk	walk while talking
mod Picasso painter	Picasso the painter
mod of examination patient	the examination of the patient

The relation between a predicate and a non clausal subject; where appropriate is obj after passivisation; for example

ncsubj arrive John _	John arrived in Paris
ncsubj employ Microsoft _	Microsoft employed C programmers
ncsubj employ Paul obj	Paul was employed by IBM

The relation between a predicate and a clausal subject con trolled from within and from without respectively; for example

csubj mean leave _	that Nellie left without saying good bye meant she was angry
csubj astonish owe _	that he owed anything would have astonished his mother
xsubj require win _	to win the America s Cup requires heaps of cash

The relation between a predicate and a direct object; where appropriate is after dative shift; e g

dobj read book _	read books
dobj mail Mary iobj	mail Mary the contract

The relation between a predicate and the second non clausal