An Iterative Approach to Estimating Frequencies over a Semantic Hierarchy

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

Abstract

This paper is concerned with using a semantic hierarchy to estimate the frequency with which a word sense appears as a given argument of a verb, assuming the data is not sense disambiguated. The standard approach is to split the count for any noun appearing in the data equally among the alternative senses of the noun. This can lead to inaccurate estimates. We describe a re-estimation process which uses the accumulated counts of hypernyms of the alternative senses in order to redistribute the count. In order to choose a hypernym for each alternative sense, we employ a novel technique which uses a χ^2 test to measure the homogeneity of sets of concepts in the hierarchy.

1 Introduction

Knowledge of the constraints a verb places on the semantic types of its arguments (variously called selectional restrictions, selectional preferences, selectional constraints) is of use in many areas of natural language processing, particularly structural disambiguation. Recent treatments of selectional restrictions have been probabilistic in nature (Resnik, 1993), (Li and Abe, 1998), (Ribas, 1995), (McCarthy, 1997), and estimation of the relevant probabilities has required corpus-based counts of the number of times word senses, or concepts, appear in the different argument positions of verbs. A difficulty arises due to the absence of a large volume of sense disambiguated data, as the counts have to be estimated from the nouns which appear in the corpus, most of which will have more than one sense. The techniques in Resnik (1993), Li and Abe (1998) and Ribas (1995) simply distribute the count equally among the alternative senses of a noun. Abney and Light (1998) have attempted to obtain selectional preferences using the Expectation Maximization algorithm by encoding WordNet as a hidden Markov model and using a modified form of the forward-backward algorithm to estimate the parameters.

The approach proposed in this paper is to use a re-estimation process which relies on counts being passed up a semantic hierarchy, from the senses of nouns appearing in the data. We make use of the semantic hierarchy in WordNet (Fellbaum, 1998), which consists of word senses, or concepts,[1] related by the 'is-a' or 'is-a-kind-of' relation. If c' is a kind of c, then c is a *hypernym* of c', and c' a *hyponym* of c. Counts for any concept are transmitted up the hierarchy to all of the concept's hypernyms. Thus if *eat chicken* appears in the corpus, the count is transmitted up to <`meat`>, <`food`>, and all the other hypernyms of that sense of *chicken*.[2] The problem is how to distinguish the correct sense of *chicken* in this case from incorrect senses such as <`wimp`>.[3] We utilise the

[1] We use the words *sense* and *concept* interchangeably to refer to a node in the semantic hierarchy.

[2] We use italics when referring to words, and angled brackets when referring to concepts or senses. This notation does not always pick out a concept uniquely, but the particular concept being referred to should be clear from the context.

[3] The example used here is adapted from McCarthy (1997). There are in fact four senses of *chicken* in WordNet 1.6, but for ease of exposition we consider only two. The hypernyms of the

fact that whilst splitting the count equally can lead to inaccurate estimates, counts do tend to accumulate in the right places. Thus counts will appear under <food>, for the object of *eat*, but not under <person>, indicating that the object position of *eat* is more strongly associated with the set of concepts dominated by <food> than with the set of concepts dominated by <person>. By choosing a hypernym for each alternative sense of *chicken* and comparing how strongly the sets dominated by these hypernyms associate with *eat*, we can give more count in subsequent iterations to the food sense of *chicken* than to the wimp sense.

A problem arises because these two senses of *chicken* each have a number of hypernyms, so which two should be compared? The chosen hypernyms have to be high enough in the hierarchy for adequate counts to have accumulated, but not so high that the alternative senses cannot be distinguished. For example, a hypernym of the food sense of *chicken* is <poultry>, and a hypernym of the wimp sense is <weakling>. However, these concepts may not be high enough in the hierarchy for the accumulated counts to indicate that *eat* is much more strongly associated with the set of concepts dominated by <poultry> than with the set dominated by <weakling>. At the other extreme, if we were to choose <entity>, which is high in the hierarchy, as the hypernym of both senses, then clearly we would have no way of distinguishing between the two senses.

We have developed a technique, using a χ^2 test, for choosing a suitable hypernym for each alternative sense. The technique is based on the observation that a chosen hypernym is too high in the hierarchy if the set consisting of the children of the hypernym is not sufficiently homogeneous with respect to the given verb and argument position. Using the previous example, <entity> is too high to represent either sense of *chicken* because the children of <entity> are not all associated in the same way with *eat*. The set consisting of the children of <meat>, however, is homogeneous with respect to the object position of *eat*, and so <meat> is not too high a level of representation. The measure of homogeneity we use is detailed in Section 5.

2 The Input Data and Semantic Hierarchy

The input data used to estimate frequencies and probabilities over the semantic hierarchy has been obtained from the shallow parser described in Briscoe and Carroll (1997). The data consists of a multiset of 'co-occurrence triples', each triple consisting of a noun lemma, verb lemma, and argument position. We refer to the data as follows: let the universe of verbs, argument positions and nouns that can appear in the input data be denoted $\mathcal{V} = \{ v_1, \ldots, v_{k_\mathcal{V}} \}$, $\mathcal{R} = \{ r_1, \ldots, r_{k_\mathcal{R}} \}$ and $\mathcal{N} = \{ n_1, \ldots, n_{k_\mathcal{N}} \}$, respectively. Note that in our treatment of selectional restrictions, we do not attempt to distinguish between alternative senses of verbs. We also assume that each instance of a noun in the data refers to one, and only one, concept.

We use the noun hypernym taxonomy of WordNet, version 1.6, as our semantic hierarchy.[4] Let $\mathcal{C} = \{ c_1, \ldots, c_{k_\mathcal{C}} \}$ be the set of concepts in WordNet. There are approximately 66,000 different concepts. A concept is represented in WordNet by a 'synonym set' (or 'synset'), which is a set of synonymous words which can be used to denote that concept. For example, the concept 'nut', as in a crazy person, is represented by the following synset: {crackpot, crank, nut, nutcase, fruitcake, screwball}. Let $\mathrm{syn}(c) \subseteq \mathcal{N}$ be the synset for the concept c, and let $\mathrm{cn}(n) = \{ c \mid n \in \mathrm{syn}(c) \}$ be the set of concepts that can be denoted by the noun n. The fact that some nouns are ambiguous means that the synsets are not necessarily disjoint.

food sense are <poultry>, <bird>, <meat>, <foodstuff>, <food>, <substance>, <object>, <entity>. The hypernyms of the wimp sense are <weakling>, <person>, <life_form>, <entity>.

[4]There are other taxonomies in WordNet, but we only use the noun taxonomy. Hence, from now on, when we talk of concepts in WordNet, we mean concepts in the noun taxonomy only.

The hierarchy has the structure of a directed acyclic graph,[5] with the isa $\subseteq \mathcal{C} \times \mathcal{C}$ relation connecting nodes in the graph, where $(c', c) \in$ isa implies c' is a kind of c. Let isa$^* \subseteq \mathcal{C} \times \mathcal{C}$ be the transitive, reflexive closure of isa, and let

$$\overline{c} = \{\, c' \mid (c', c) \in \text{isa}^* \,\}$$

be the set consisting of the concept c and all of its hyponyms. The set $\overline{\langle\texttt{food}\rangle}$ contains all the concepts which are kinds of food, including <food>.

Note that words in our data can appear in synsets anywhere in the hierarchy. Even concepts such as <entity>, which appear near the root of the hierarchy, have synsets containing words which may appear in the data. The synset for <entity> is {entity, something}, and the words *entity* and *something* may well appear in the argument positions of verbs in the corpus. Furthermore, for a concept c, we distinguish between the set of words that can be used to denote c (the synset of c), and the set of words that can be used to denote concepts in $\overline{c}$.[6]

3 The Measure of Association

We measure the association between argument positions of verbs and sets of concepts using the **association norm** (Abe and Li, 1996).[7] For $C \subseteq \mathcal{C}$, $v \in \mathcal{V}$ and $r \in \mathcal{R}$, the association norm is defined as follows:

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

For example, the association between the object position of *eat* and the set of concepts denoting kinds of food is expressed as follows: $\mathrm{A}(\overline{\langle\texttt{food}\rangle}, \mathit{eat}, \text{object})$. Note that, for $C \subseteq \mathcal{C}$, $p(C|v, r)$ is just the probability of the disjunction of the concepts in C; that is,

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

In order to see how $p(c|v, r)$ relates to the input data, note that given a concept c, verb v and argument position r, a noun can be generated according to the distribution $p(n|c, v, r)$, where

$$\sum_{n \in \mathrm{syn}(c)} p(n|c, v, r) = 1$$

Now we have a model for the input data:

$$\begin{aligned} p(n, v, r) &= p(v, r) p(n|v, r) \\ &= p(v, r) \sum_{c \in \mathrm{cn}(n)} p(c|v, r) p(n|c, v, r) \end{aligned}$$

Note that for $c \notin \mathrm{cn}(n)$, $p(n|c, v, r) = 0$.

The association norm (and similar measures such as the mutual information score) have been criticised (Dunning, 1993) because these scores can be greatly over-estimated when frequency counts are low. This problem is overcome to some extent in the scheme presented below since, generally speaking, we only calculate the association norms for concepts that have accumulated a significant count.

The association norm can be estimated using maximum likelihood estimates of the probabilities as follows.

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

4 Estimating Frequencies

Let freq(c, v, r), for a particular c, v and r, be the number of (n, v, r) triples in the data in which n is being used to denote c, and let freq(v, r) be the number of times verb v appears with something in position r in the data; then the relevant maximum likelihood estimates, for $c \in \mathcal{C}$, $v \in \mathcal{V}$, $r \in \mathcal{R}$, are as

[5] The number of nodes in the graph with more than one parent is only around one percent of the total.

[6] Note that Resnik (1993) uses rather non-standard terminology by refering to this second set as the synsets of c.

[7] This work restricts itself to verbs, but can be extended to other kinds of predicates that take nouns as arguments, such as adjectives.

follows.

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

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

Since we do not have sense disambiguated data, we cannot obtain $\mathrm{freq}(c,v,r)$ by simply counting senses. The standard approach is to estimate $\mathrm{freq}(c,v,r)$ by distributing the count for each noun n in $\mathrm{syn}(c)$ evenly among all senses of the noun as follows:

$$\hat{\mathrm{freq}}(c,v,r) = \sum_{n \in \mathrm{syn}(c)} \frac{\mathrm{freq}(n,v,r)}{|\,\mathrm{cn}(n)|}$$

where $\mathrm{freq}(n,v,r)$ is the number times the triple (n,v,r) appears in the data, and $|\,\mathrm{cn}(n)|$ is the cardinality of $\mathrm{cn}(n)$.

Although this approach can give inaccurate estimates, the counts given to the incorrect senses will disperse randomly throughout the hierarchy as noise, and by accumulating counts up the hierarchy we will tend to gather counts from the correct senses of related words (Yarowsky, 1992; Resnik, 1993). To see why, consider two instances of possible triples in the data, *drink wine* and *drink water*. (This example is adapted from Resnik (1993).) The word *water* is a member of seven synsets in WordNet 1.6, and *wine* is a member of two synsets. Thus each sense of *water* will be incremented by 0.14 counts, and each sense of *wine* will be incremented by 0.5 counts. Now although the incorrect senses of these words will receive counts, those concepts in the hierarchy which dominate more than one of the senses, such as <`beverage`>, will accumulate more substantial counts.

However, although counts tend to accumulate in the right places, counts can be greatly underestimated. In the previous example, $\hat{\mathrm{freq}}(\overline{\texttt{<beverage>}},\mathit{drink},\mathrm{object})$ is incremented by only 0.64 counts from the two data instances, rather than the correct value of 2.

The approach explored here is to use the accumulated counts in the following re-estimation procedure. Given some verb v and position r, for each concept c we have the following initial estimate, in which the counts for a noun are distributed evenly among all of its senses:

$$\hat{\mathrm{freq}}^{0}(c,v,r) = \sum_{n \in \mathrm{syn}(c)} \frac{\mathrm{freq}(n,v,r)}{|\,\mathrm{cn}(n)|}$$

Given the assumption that counts from the related senses of words that can fill position r of verb v will accumulate at hypernyms of c, let $\mathrm{top}(c,v,r)$ be the hypernym of c (or possibly c itself) that most accurately represents this set of related senses. In other words, $\overline{\mathrm{top}(c,v,r)}$ will be an approximation of the set of concepts related to c that fill position r of verb v. Rather than splitting the counts for a noun n evenly among each of its senses $c \in \mathrm{cn}(n)$, we distribute the counts for n on the basis of the accumulated counts at $\mathrm{top}(c,v,r)$ for each $c \in \mathrm{cn}(n)$. In the next section we discuss a method for finding $\mathrm{top}(c,v,r)$, but first we complete the description of how the re-estimation process uses the accumulated counts at $\mathrm{top}(c,v,r)$.

Given a concept c, verb v and position r, in the following formula we use $\overline{[c,v,r]}$ to denote the set of concepts $\overline{\mathrm{top}(c,v,r)}$. The re-estimated frequency $\hat{\mathrm{freq}}^{m+1}(c,v,r)$ is given as follows.

$$\hat{\mathrm{freq}}^{m+1}(c,v,r) = \sum_{n \in \mathrm{syn}(c)} \mathrm{freq}(n,v,r) \frac{\hat{\mathrm{A}}^{m}(\overline{[c,v,r]},v,r)}{\sum_{c' \in \mathrm{cn}(n)} \hat{\mathrm{A}}^{m}(\overline{[c',v,r]},v,r)}$$

Note that only nouns n in $\mathrm{syn}(c)$ contribute to the count for c. The count $\mathrm{freq}(n,v,r)$ is split among all concepts in

c	$\widehat{\mathrm{freq}}^0(\overline{c},eat,\mathrm{obj})$	$\widehat{\mathrm{freq}}^0(\overline{c},\mathrm{obj})-$ $\widehat{\mathrm{freq}}^0(\overline{c},eat,\mathrm{obj})$	$\widehat{\mathrm{freq}}^0(\overline{c},\mathrm{obj}) =$ $\sum_{v\in\mathcal{V}} \widehat{\mathrm{freq}}^0(\overline{c},v,\mathrm{obj})$
`<milk>`	0.0 (0.6)	9.0 (8.4)	9.0
`<meal>`	8.5 (5.6)	78.0 (80.9)	86.5
`<course>`	1.3 (1.7)	24.7 (24.3)	26.0
`<dish>`	5.3 (5.7)	82.3 (81.9)	87.6
`<delicacy>`	0.3 (1.8)	27.4 (25.9)	27.7
	15.4	221.4	236.8

Table 1: Contingency table for children of `<nutriment>`

cn(n) according to the ratio

$$\frac{\hat{\mathrm{A}}^m(\overline{[c,v,r]},v,r)}{\sum_{c'\in\mathrm{cn}(n)} \hat{\mathrm{A}}^m(\overline{[c',v,r]},v,r)}$$

For a set of concepts C,

$$\hat{\mathrm{A}}^m(C,v,r) = \frac{\hat{p}^m(C|v,r)}{\hat{p}^m(C|r)}$$

where

$$\hat{p}^m(C|v,r) = \frac{\widehat{\mathrm{freq}}^m(C,v,r)}{\widehat{\mathrm{freq}}(v,r)}$$

$$\hat{p}^m(C|r) = \frac{\sum_{v\in\mathcal{V}} \widehat{\mathrm{freq}}^m(C,v,r)}{\sum_{v\in\mathcal{V}} \widehat{\mathrm{freq}}(v,r)}$$

$$\widehat{\mathrm{freq}}^m(C,v,r) = \sum_{c\in C} \widehat{\mathrm{freq}}^m(c,v,r)$$

5 Determining top(c,v,r)

The technique for calculating top(c,v,r) is based on the assumption that a hypernym c' of c is too high in the hierarchy to be top(c,v,r) if the children of c' are not sufficiently homogeneous with respect to v and r. A set of concepts, C, is taken to be homogeneous with respect to a given $v \in \mathcal{V}$ and $r \in \mathcal{R}$, if $p(v|\overline{c},r)$ has a similar value for each $c \in C$. Note that this is equivalent to comparing association norms since

$$\begin{aligned} p(v|C,r) &= \frac{p(C|v,r)}{p(C|r)}p(v|r) \\ &= \mathrm{A}(c,v,r)p(v|r) \end{aligned}$$

and, as we are considering homogeneity for a given verb and argument position, $p(v|r)$ is a constant.

To determine whether a set of concepts is homogeneous, we apply a χ^2 test to a contingency table of frequency counts. Table 1 shows frequencies for the children of `<nutriment>` in the object position of *eat*, and the figures in brackets are the expected values, based on the marginal totals in the table.

Notice that we use the $\widehat{\mathrm{freq}}_0$ counts in the table. A more precise method, that we intend to explore, would involve creating a new table for each $\widehat{\mathrm{freq}}_m$, $m \geq 0$, and recalculating top(c,v,r) after each iteration. A more significant problem of this approach is that by considering $p(v|\overline{c},r)$, we are not taking into account the possibility that some concepts are associated with more verbs than others. In further work, we plan to consider alternative ways of comparing levels of association.

The null hypothesis of the test is that $p(v|\overline{c},r)$ is the same for each c in the table. For example, in Table 1 the null hypothesis is that for every concept c that is a child of `<nutriment>`, the probability of some concept $c' \in \overline{c}$ being eaten, given that it is the object of some verb, is the same. For the experiments described in Section 6, we used 0.05 as the level of significance. Further work will investigate the effect that different levels of significance have on the estimated frequencies.

The χ^2 statistic corresponding to Table 1

(v, c)	Hypernyms of c
(*eat*,<hotdog>)	<sandwich> <snack_food> ... <NUTRIMENT> <food> <substance> <entity>
(*drink*,<coffee>)	<BEVERAGE> <food> <substance> <entity>
(*see*,<movie>)	<SHOW> <communication> <social_relation> <relation> <abstraction>
(*hear*,<speaker>)	<communicator> <person> <life_form> <ENTITY>
(*kiss*,<Socrates>)	<philosopher> <intellect> <person> <LIFE_FORM> <entity>

Table 2: How log-likelihood χ^2 chooses $\mathrm{top}(c, v, r)$

is 4.8. We use the log-likelihood χ^2 statistic, rather than the Pearson's χ^2 statistic, as this is thought to be more appropriate when the counts in the contingency table are low (Dunning, 1993).[8] For a significance level of 0.05, with 4 degrees of freedom, the critical value is 9.49 (Howell, 1997). Thus in this case, the null hypothesis (that the children of <nutriment> are homogeneous with respect to eat) would not be rejected.

Given a verb v and position r, we compute $\mathrm{top}(c, v, r)$ for each c by determining the homogeneity of the children of the hypernyms of c. Initially, we let $\mathrm{top}(c, v, r)$ be the concept c itself. We work from c up the hierarchy reassigning $\mathrm{top}(c, v, r)$ to be successive hypernyms of c until we reach a hypernym whose children are not sufficiently homogeneous. In situations where a concept has more than one parent, we consider the parent which results in the lowest χ^2 value as this indicates the highest level of homogeneity.

6 Experimental Results

In order to evaluate the re-estimation procedure, we took triples from approximately two million words of parsed text from the BNC corpus using the shallow parser developed by Briscoe and Carroll (1997). For this work we only considered triples for which $r = \mathrm{obj}$. Table 2 shows some examples of how the log-likelihood χ^2 test chooses $\mathrm{top}(c, v, r)$ for various $v \in \mathcal{V}$ and $c \in \mathcal{C}$.[9] In giving the list of hypernyms the selected concept $\mathrm{top}(c, v, \mathrm{obj})$ is shown in upper case.

Table 3 shows how frequency estimates change, during the re-estimation process, for various $v \in \mathcal{V}$, $c \in \mathcal{C}$, and $r = \mathrm{obj}$. The figures in Table 3 show that the estimates appear to be converging after around 10 iterations. The first column gives the frequency estimates using the technique of splitting the count equally among alternative senses of a noun appearing in the data. The figures for *eat* and *drink* suggest that these initial estimates can be greatly underestimated (and also overestimated for cases where the argument strongly violates the selectional preferences of the verb, such as *eat* <location>). The final column gives an upper bound on the re-estimated frequencies. It shows how many nouns in the data, in the object position of the given verb, that could possibly be denoting one of the concepts in $\overline{c}$, for each v and $\overline{c}$ in the table. For example, 95 is the number of times a noun which could possibly

[8] Low counts tend to occur in the table when the test is being applied to a set of concepts near the foot of the hierarchy. A further extension of this work will be to use Fisher's exact test for the tables with particularly low counts.

[9] Notice that < hotdog > is classified at the <nutriment> level rather than <food>. This is presumably due to the fact that beverage is classed as a food, making the set of concepts $\overline{\texttt{<food>}}$ heterogenous with respect to the object position of *eat*.

$(v, \overline{c})$	$\hat{\mathrm{freq}}^{m}(\overline{c}, v, \mathrm{obj})$				Limit
	$m = 0$	$m = 1$	$m = 5$	$m = 10$	
$(\mathit{eat}, \overline{\texttt{<food>}})$	60.8	85.0	89.6	89.8	95
$(\mathit{drink}, \overline{\texttt{<beverage>}})$	10.5	22.7	23.5	23.4	26
$(\mathit{eat}, \overline{\texttt{<location>}})$	2.0	1.2	1.1	1.1	6
$(\mathit{see}, \overline{\texttt{<obj>}})$	237.1	235.7	240.2	240.3	568
$(\mathit{hear}, \overline{\texttt{<person>}})$	90.8	85.5	85.5	85.5	130
$(\mathit{enjoy}, \overline{\texttt{<amusement>}})$	2.9	3.1	3.3	3.3	5
$(\mathit{measure}, \overline{\texttt{<abstraction>}})$	19.1	21.7	23.3	23.4	31

Table 3: Example of re-estimated frequencies

be denoting a concept dominated by `<food>` appeared in the object position of *eat*. Since *eat* selects so strongly for its object, we would expect freq(`<food>`,*eat*,obj) (i.e., the true figure) to be close to 95. Similarly, since *drink* selects so strongly for its object, we would expect freq(`< beverage >`,*drink*,obj) to be close to 26. We would also expect freq(`<location>`,*eat*,obj) to be close to 0. As can be seen from Table 3, our estimates converge quite closely to these values.

It is noticeable that the frequency counts for weakly selecting verbs do not change as much as for strongly selecting verbs. Thus, the benefit we achieve compared to the standard approach of distributing counts evenly is reduced in these cases. In order to investigate the extent to which our technique may be helping, for each triple in the data we calculated how the distribution of the count changed due to our re-estimation technique. We estimated the extent to which the distribution had changed by calculating the percentage increase in the count for the most favoured sense after 10 iterations. Table 4 shows the results we obtained. The proportions given in the second column are of the triples in the data containing nouns with more than one sense.[10] We can see from the table that for 43% of the triples our technique is having little effect, but for 23% the count is at least doubled.

[10] 17% of the data involved nouns with only one sense in WordNet.

7 Conclusions

We have shown that the standard technique for estimating frequencies over a semantic hierarchy can lead to inaccurate estimates. We have described a re-estimation procedure which uses an existing measure of selectional preference and which employs a novel way of selecting a hypernym of a concept. Our experiments indicate that the re-estimation procedure gives more accurate estimates than the standard technique, particularly for strongly selecting verbs. This could prove particularly useful when using selectional restrictions, for example in structural disambiguation.

8 Acknowledgements

This work is supported by UK EPSRC project GR/K97400 and by an EPSRC Research Studentship to the first author. We would like to thank John Carroll for supplying and providing help with the data, and Diana McCarthy, Gerald Gazdar, Bill Keller and the anonymous reviewers for their helpful comments.

Percentage Increase	Proportion of data
0–10	43%
10–50	18%
50–100	16%
100–	23%

Table 4: How the distribution of counts change

References

Naoki Abe and Hang Li. 1996. Learning Word Association Norms using Tree Cut Pair Models. In *Proceedings of the Thirteenth International Conference on Machine Learning.*

Steve Abney and Marc Light. 1998. Hiding a Semantic Class Hierarchy in a Markov Model. Unpublished. Paper can be obtained from http://www.ims.uni-stuttgart.de/~light/onlinepapers.html.

Ted Briscoe and John Carroll. 1997. Automatic extraction of subcategorization from corpora. In *Proceedings of the 5th ACL Conference on Applied Natural Language Processing*, pages 356–363, Washington, DC.

Ted Dunning. 1993. Accurate Methods for the Statistics of Surprise and Coincidence. *Computational Linguistics*, 19(1):61–74.

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

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

Hang Li and Naoki Abe. 1998. Generalizing Case Frames using a Thesaurus and the MDL Principle. *Computational Linguistics*, 24(2):217–244.

Diana McCarthy. 1997. Word sense disambiguation for acquisition of selectional preferences. In *Proceedings of the Proceedings of the ACL/EACL 97 Workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*, pages 52–61, Madrid, Spain.

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

Francesc Ribas. 1995. On Learning More Appropriate Selectional Restrictions. In *Proceedings of the Seventh Conference of the European Chapter of the Association for Computational Linguistics*, Dublin, Ireland.

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